nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

CELLULAR POWER GRID

Mitochondrial network conduction distributes energy in muscle cells **PAGE 617**

STEM CELLS

MARCH OF THE ORGANOIDS

The proliferation of lab-grown mini-organs

PAGE 520

HISTORY

AFTER HIROSHIMA

The true legacy of the 'physicists' war'

PAGE 523

REVIEWS

SUMMER BOOKS

Holiday reads from Nature's reviewers and editors

PAGE 528

➲ NATURE.COM/NATURE

30 July 2015 £10

Vol. 523, No. 7562

31

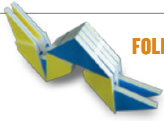
9 770028 083095

THIS WEEK

EDITORIALS

DISEASE We need to get over our outbreak overreactions **p.502**

WORLD VIEW How the Pope could end the culture wars **p.503**



FOLDING Chunky origami technique could make solar panels **p.505**

Tropical protection

After years of talk, the palm-oil industry is looking into adopting environmental standards. Such rules must be strong, and need to be implemented.

More than 100 major companies worldwide have made commitments to promote the use of environmentally sustainable palm oil over the past few years. This is to their credit. Palm oil finds its way into everything from food and cosmetics to biofuels, but the expansion of palm plantations has driven widespread deforestation — as well as carbon emissions — in places such as Indonesia and Malaysia. To various degrees, companies that trade in palm oil have promised to halt the use of oil from newly cleared land, but implementing such goals is not easy. The latest attempt to create workable standards comes from an industry consortium in consultation with a team of respected scientists. Their report is due out in December, and a draft is available for public comment until 31 July (see go.nature.com/rt7fue).

This High Carbon Stock (HCS) Study was formally launched last year, when five leading palm-oil producers, including Sime Darby in Kuala Lumpur and IOI Corporation Berhad in Putrajaya, Malaysia, signed the Sustainable Palm Oil Manifesto. That document commits signatories to halting the expansion of palm plantations in dense forests where carbon emissions would be highest, but says that the palm-oil industry cannot focus solely on environmental issues. Environmentalists immediately accused the companies of seeking to undermine attempts to produce a stricter set of guidelines, and to delay obvious solutions with complicated science.

There is some truth to this, but the merits of a given project do depend in part on the social and economic context in which it is situated. Decisions about land use are rarely made on the basis of environmental criteria alone, and many of the regions in which the plantations are located — or will be located — would see social and economic benefits from an orderly palm-oil industry.

The question is where to draw the line. Most would agree that it does not make sense to tear down old-growth forests, which store a lot of carbon and are home to a diverse array of plants and animals. The same could probably be said for selectively logged forests, where only the biggest and most valuable trees have been taken, which are still high in carbon and biodiversity. Everybody agrees that it would be wise to focus development on abandoned land that has already been fully cleared, and so has little carbon or biodiversity to speak of; in such areas, a palm plantation could increase the carbon stock, thereby alleviating global warming. In between, on degraded and heavily logged forests and in areas where forests are actively regrowing, there is more room for debate.

The current draft of the HCS Study report seeks to create a framework for evaluating projects on the basis of both land type and socio-economic conditions. It proposes classifying land according to the state of forests: at the extremes, green represents the go-zone, such as already cleared land, and red the no-go zone, where primary forest remains. In the centre is ambiguous amber, a middle zone in which trade-offs are possible. If the social and economic benefits are high enough, perhaps a small hit to the climate is acceptable and could be offset by protecting additional land elsewhere. The first step in making such decisions is to get data on forest cover, and the study advocates mapping land with both high-resolution satellites and aircraft-based lasers to gather detailed measurements of forest structure.

Confusingly, before the HCS Study launched, major environmental groups were engaging the industry in separate negotiations known as the High Carbon Stock Approach. Those talks intended to create a more conservative set of guidelines that often default to the red no-go zone when it comes to development. The HCS Study consciously goes in the other direction, acknowledging that there may be cases in which natural forests could be converted to plantations in the name of alleviating poverty. "This is the essence of the 'quid pro quo' explored in this Study," the authors write.

"The industry must find a way to promote both environmental protection and social well-being."

Ultimately, the industry must to find a way to promote both environmental protection and social well-being. Finding the right formula will not be easy, but it is a sign of progress that all sides are seeking a solution. In theory, this is the duty of government, but governments across the tropics have had a hard time controlling rampant development that has left many citizens behind. It would be a step in the right direction for environmentalists, scientists and businesses to agree on a set of meaningful standards. Then it would be a matter of ensuring that companies keep their word. ■

Secret service

Government labs should be subject to the same transparent oversight as academic facilities.

The 'overabundance of caution' used by national defence and security agencies can border on the ridiculous. US government paranoia over terrorism led to the generally despised — and questionably effective — airport rituals of prohibiting bottles that contain more than 100 millilitres of most liquids and subjecting all passengers to radiation in a virtual strip search. Public panic led to similarly overblown US responses to the 2014 Ebola outbreak, including the forced quarantine of people who were never exposed to the virus and had no chance of causing an epidemic (see page 502).

How, then, was the US Department of Defense (DOD) able this year to send live anthrax spores across at least seven international borders and to at least 183 labs without the authorities noticing? If there is anywhere that paranoid officials should want to monitor when it comes to anthrax, it is the DOD. After all, the DOD works with more anthrax than any other institution, and the only known bioterror attack using anthrax spores as a weapon originated at a DOD lab.

Oversight systems seem to have been watching everything except the most likely source of a threat.

When this year's failure came to light, the DOD immediately began a 30-day investigation of itself. Its 38-page conclusion, released to the public last week, blamed no one in particular (see go.nature.com/ltcn6f). The military determined that the radiation procedure being used at the lab — Dugway Proving Ground in Utah — to kill the spores was ineffective. It emphasizes that no one was harmed, and that there is no proven method to kill the notoriously resilient spores. Both these things are true.

What is still unclear, however, is why the procedure was not better tested. The US Centers for Disease Control and Prevention (CDC) does not have particular standards for inactivation protocols. But if it did, Dugway's protocol surely would not meet them: the lab had never optimized the procedure, and the base's own records showed that the process failed once in every five attempts. Furthermore, neither the sending nor the receiving labs had done enough to verify that the samples were dead. Dugway, for instance, tested only 5% of each sample for viability, which would not have detected a low concentration of live spores. In a twist of irony, DOD scientist Bruce Ivins, who was allegedly responsible for the 2001 anthrax attacks, had suggested that half of a sample should be screened to rule out viability.

Dugway has been in hot water before. An investigation by the news outlet *USA Today* found that the CDC had reprimanded the facility eight years ago for using a different experimental protocol to inactivate anthrax spores and then shipping them even when tests showed that they were still alive. According to *USA Today*, Dugway was let off with a warning, and the incident was not included in the DOD's annual report to Congress.

Academic labs could be justifiably rankled at the amount of money and time they have to spend complying with regulations on less dangerous pathogens and harmless amounts of radiation. A university that flouts CDC regulations would probably be subject to harsh penalties. But US law allows government labs to maintain secrecy around their procedures and the results of investigations into their biosafety mishaps, of which there seem to be many.

"It should not be left up to the media to discover serious accidents at agencies."

That could soon change. On 28 July, both the DOD and the CDC were hauled before a congressional committee that is demanding answers and a new probe into the latest incident. The committee has also called for the agencies to produce a list of the labs that are authorized to work with anthrax and other bioterror agents, and for details of biosafety violations. Earlier this month, the CDC announced that it is beginning a 90-day review of its biosafety procedures for federal research labs that work with dangerous pathogens.

It should not be left up to the media to discover serious accidents at the agencies charged with protecting people from bioterrorism. To be clear, the research they perform on anthrax and other pathogens is essential for biosecurity. Incompetent oversight combined with a culture of secrecy could threaten that work. And, given the overabundance of caution applied elsewhere, there should be some spare to deploy at the government labs at which it is most needed. ■

Realistic risks

The communication of risk in disease outbreaks is too often neglected; that must change.

The outbreak of Middle East respiratory syndrome (MERS) in South Korean hospitals is effectively over, with no new cases since 2 July. Since it began on 11 May, a total of just 186 people were infected by the coronavirus, 36 of whom have died. The episode was tragic, but its economic and social impact was disproportionate. If the world is to respond effectively to infectious-disease outbreaks, then the authorities, the media and communities must pay more attention to risk communication.

The only people at real risk of infection in South Korea were those who had shared a hospital area with someone who had MERS. Yet at the outbreak's peak in early June, thousands of schools were needlessly closed and public events were cancelled. Tourist numbers dropped by 41% compared with the same month last year: a US$10-billion loss that is expected to knock 0.1% off the country's gross domestic product growth this year. The only winners were those selling the ubiquitous and superfluous face masks.

One important question — and lesson to learn — is how the authorities failed both to convey the limited threat posed by MERS, and to persuade the media and public that they had the outbreak under control.

Public trust in Korean officials was already low after a perceived bungled response to the sinking of the ferry MV *Sewol* last year, which killed more than 300 people, many of them secondary-school pupils. When MERS struck, the authorities foolishly declined to identify the affected hospitals publicly, allowing rumours — amplified by social media — to fill the space. This faltering start was unfortunate because the government did get its act together soon after. Its transparency in reporting new cases became exemplary, as did its public-health response — including the massive task of tracing and isolating the more than 16,500 people who had been in contact with infected patients. The last contact was released from isolation this week.

Disease outbreaks are frightening, and overreaction to a virus that can kill is an understandable human response. It is one that needs to be understood and managed, not dismissed as irrational.

This puts great responsibility on the shoulders of the press and politicians, and often we see that some are not up to the job. When a handful of Ebola cases occurred on US soil last year, it sparked what President Barack Obama has described as "hysteria". Many media reports were balanced and excellent, but too much of the reporting was excessive and sensationalist. Complicating matters further, right-wing political opportunists and pundits used the Ebola cases to take partisan shots at the Obama administration. Combined with the 24/7 news cycle, and again amplified by social media, coverage of what was a legitimate news story became a shambolic and sorry mess, utterly detached from the reality — that the United States faced no threat of an Ebola epidemic.

This had real consequences. Several politicians, including Chris Christie, the governor of New Jersey, implemented unnecessary and counterproductive measures, such as forced quarantine of US health-care workers returning from West Africa. Republican presidential hopeful Donald Trump showed a troubling grasp of the issue, and called for US borders to be sealed to those arriving from the region, including health-care workers. If this was the US response to a non-existent disease threat, what would its reaction be to a serious epidemic threat? Some outbreak-response officials think that the trend towards instantaneous news, compounded by social media, could interfere with effective public-health interventions and result in societal chaos.

Overreactions to outbreaks that pose no large threat can distract from those that do, and the priority is to eliminate the threats at source. Ebola must be stamped out in West Africa, and MERS must not be allowed to fester in the Middle East, where it is endemic in camels. Researchers need to identify and close the routes by which the MERS virus spreads to people. Social-science researchers can help to unravel complex factors affecting public reactions to outbreaks, and how authorities can build trust, so that risks can be better communicated. They might also ask how European countries managed to respond coolly to the arrival of both MERS and Ebola cases. ■

➲ **NATURE.COM**
To comment online, click on Editorials at:
go.nature.com/xhunqv

UNIV. NOTRE DAME

Faith and science can find common ground

Pope Francis has found a meeting place for those with extreme religious and environmentalist stances, says **David M. Lodge***.*

In recent weeks, we have learned that Pope Francis enticed Cuban President Raúl Castro to consider a return to Catholicism, and has ended a dispute involving US nuns that will allow them to return to serving the poor free from the suspicion of heresy.

Perhaps most surprisingly, at least to this Protestant ecologist embedded for 30 years in a Roman Catholic university, the Pope has suggested that humans should not breed "like rabbits", despite his church's continued prohibition of birth control.

Pope Francis is clearly a man on a mission to shake things up. Could the world's leading Catholic help to bridge the divide between science and the Protestant views that dominate the religious 'anti-science' movement? I think that he could.

In his recent encyclical on humans and the environment, Pope Francis described environmental degradation with great scientific accuracy, and he linked it to economic exploitation and the plight of the poor. This is a challenge to many conservative Protestants who believe that humans, because they are made in God's image, have a divine right to exploit the natural world.

The Pope's argument is a powerful one, and addresses those with power, especially in the United States. The views of notorious climate-sceptic Senator James Inhofe (Republican, Oklahoma) on the threat of global warming, for example, are underpinned by his strong Protestant conviction that God created natural resources for humans, and that we are arrogant in thinking that we can affect God's plan for the Earth. By contrast, many environmentalists argue that humankind should protect nature for its intrinsic value, with little apparent regard for the importance of its use for human welfare.

IS IT SO **SURPRISING** THAT, IN THE UNITED STATES ESPECIALLY, **ATHEISM** IS OVER-REPRESENTED AMONG **SCIENTISTS?**

By framing protection of the environment as protecting human welfare, the Pope has linked the interests of groups that are often at odds. He offers some middle ground on which both sides of this polarized debate can meet and work towards a mutually desirable future.

Such a compromise between the extremes of the religious and environmentalist positions could also help to defuse other sources of tension between faith and science. To many people, the two cannot be reconciled — so much so that when I tell people I am a biologist, believe in evolution and work on environmental issues, I am often told that I cannot be a Christian. Sadly, this is the message in many conservative Protestant churches: choose between science and faith.

The same polarization is urged by many prominent popularizers of science and the 'New Atheists' — with Richard Dawkins as their figurehead. Is it so surprising, then, that in the United States especially, atheism is over-represented among scientists, and that science–faith polarization is increasingly reflected in political and cultural discourse?

➲ NATURE.COM
Discuss this article online at:
go.nature.com/rgzfi6

For example, nothing in the official teaching of Catholicism opposes evolution. Creationism is a recent Protestant invention, based on extreme, literal interpretations of the first three chapters of the Bible's book of Genesis. Catholicism relies more on an interpretation of the scriptures that is rooted in a tradition of reason informing faith. Yet when I ask my biology undergraduates whether they feel a conflict between their faith and evolution, about half of every class — 85% of whom are Catholic — say yes.

The students respond in this way because evolution, alongside issues such as climate change, stem cells, abortion and gay marriage, has been conscripted into the culture wars, with science increasingly suffering collateral damage. And as the culture wars have forced people to choose sides, respect for science is now divided along political lines too, with huge influence on policy.

US environmental-protection policies, which began as bipartisan efforts to protect human and environmental health, have become destructively partisan. It was the Republican president Richard Nixon who in 1972 signed the Clean Water Act that brought Lake Erie back from the dead, restoring one of the most economically valuable freshwater fisheries in the world. Nowadays, efforts to improve water and air quality are too often supported by Democrats and opposed by Republicans — on the grounds that environmental protection harms human welfare, and that because the world is temporary, long-term protection is unnecessary. They dismiss scientists, who increasingly quantify the great extent to which environmental protection benefits humans, as just another special-interest group.

As a Protestant scientist, I am distressed to see my faith twisted into support for such short-sighted extremism. Martin Luther, the great Protestant reformer, once said: "Even if I knew that tomorrow the world would go to pieces, I would still plant my apple tree." Like Pope Francis, he understood the importance of loving and tending the gift of creation.

If Pope Francis can persuade the communist Raúl Castro to reconsider Catholicism, I can hope that the Pope's respect for the scientific consensus on climate change will foster a more constructive dialogue between the communities of science, faith and policymakers. His recognition that the economy and the environment are inextricably linked, especially for the desperately poor, builds on a foundation that is older and deeper than the recent US culture wars. ■

David M. Lodge *is director of the University of Notre Dame Environmental Change Initiative, Indiana, and editor (with historian Christopher Hamlin) of* Religion and the New Ecology *(2006).*
e-mail: dlodge@nd.edu

RESEARCH HIGHLIGHTS

Selections from the scientific literature

QUANTUM INFORMATION

Qubit control in a 3D matrix

Qubits — quantum bits which store and process information in quantum computers — can be controlled individually in a 3D structure without disturbing nearby atoms.

Neutral atoms show promise as qubits when they are cooled and trapped by light, but manipulating one atom without disturbing its neighbours is difficult. David Weiss and his team at Pennsylvania State University in University Park controlled a single atom in a 5 × 5 × 5 array of trapped caesium atoms by firing two beams of circularly polarized light so that they intersected at the target atom. This caused the energy levels of electrons in the atom to shift, allowing the researchers to change its quantum state by hitting it with microwaves.

The method should make it easier to scale up quantum computers that use this kind of qubit, the researchers say.

Phys. Rev. Lett. 115, **043003 (2015)**

SCOTT M. BOBACK

ANIMAL BEHAVIOUR

How boa constrictors really kill

Animals that have been squeezed to death by snakes probably die of circulatory arrest rather than of suffocation as was thought.

Scott Boback at Dickinson College in Carlisle, Pennsylvania, and his team anaesthetized rats and implanted probes and catheters to measure their heart rate, blood pressure and blood chemistry as the animals were being squeezed by a boa (*Boa constrictor*; pictured). Constriction lasted an average of 6.5 minutes, but the rats' peripheral blood pressure dropped by half as early as 6 seconds in. By the end of constriction, the rats' blood chemistry showed signs of system-wide circulatory problems and there was evidence that the heart had undergone significant electrical dysfunction.

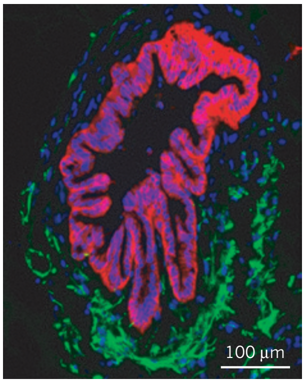
The authors suggest that snakes may release their prey only after detecting that it has experienced irreversible heart failure.

J. Exp. Biol. 218, **2279–2288 (2015)**

STEM CELLS

Stomach tissue made in a dish

Mouse embryonic stem cells can develop into 3D 'mini stomachs' in the lab.



TAKA-AKI K. NOGUCHI *ET AL.*/*NATURE CELL BIOL.*

Researchers have previously used stem cells to make parts of the stomach, but not the whole organ. Taka-aki Noguchi, Akira Kurisaki at the University of Tsukuba in Japan and their team made stomach tissue — including the main food-containing part — by adding several key growth factors to the stem-cell culture after six days. These turned on expression of the *Barx1* gene, which is essential for stomach development. After about 60 days, the cells developed into stomach tissue (**pictured**) that contained several specialized types of stomach cell. The mini stomach also secreted a digestive hormone and acid, and had a similar gene-expression profile to that of adult stomach tissue.

This system could be used to study stomach diseases, the authors say.

Nature Cell Biol. **http://doi.org/6gf (2015)**

For a related News Feature on organoids, see page 520.

NEUROBIOLOGY

Marijuana's good without the bad

Mice treated with marijuana's active component, THC, along with other key molecules, can experience its pain-relieving benefits without the usual memory impairments.

THC binds to the CB1 cannabinoid receptor in the brain, causing negative effects such as poor memory and anxiety. However, it also affects behaviours that are regulated by the serotonin 2A receptor. Patricia Robledo at Pompeu Fabra University in Barcelona, Spain, and her colleagues gave THC to mice lacking the serotonin receptor, and found that the animals did not show signs of memory loss but could still tolerate painful stimuli. Studies of cells expressing both types of receptors revealed that the receptors physically interact. Mice that were given THC and peptides that stop the receptors interacting in the brain did not show memory problems, but still experienced pain relief.

The findings suggest a way to minimize the negative effects of medical marijuana, the authors say.

PLoS Biol. 13, **e1002194 (2015)**

ASTRONOMY

Telescope spies early galaxy's birth

Astronomers have spotted the glow from one of the most distant galaxies ever seen in the early Universe.

Roberto Maiolino at the University of Cambridge, UK, and his colleagues used the high-resolution Atacama Large Millimeter/submillimeter Array (ALMA) telescope in Chile to observe three faint galaxies that began forming less than one billion years after the Big Bang. In one galaxy they detected clouds of cold ionized carbon that was shifted away from the bright, star-forming centre. This matches models of early galaxy formation, which predict that active young stars disperse such clouds.

The data will help to test theories about how the Universe's first stars and galaxies formed, the team says.

Mon. Not. R. Astron. Soc. 452, **54–68 (2015)**

NEUROBIOLOGY

A critical period for brain health

Disrupting a signalling molecule in week-old mice prevents their neurons from altering connections with other neurons in adulthood.

Cognitive problems can result from an inability to strengthen or weaken these connections, called synapses, in the brain in response to external triggers — a process known as plasticity. Neil Hardingham of Cardiff University, UK, and his colleagues blocked signalling by the DISC1 protein in 7-day-old mice for 6–48 hours. When the animals reached adulthood, their synapses failed to adapt in response to signals generated by stimulating one of their whiskers.

In humans, changes to DISC1 signals during this critical period soon after birth could underlie psychiatric symptoms that emerge during adolescence, the authors suggest.

Science 349, **424–427 (2015)**

MICROBIOLOGY

Microbe war waged with biofilms

Sticky films of microbes form as the result of competition, not cooperation, between bacterial strains.

Scientists had thought that biofilms form cooperatively, with multiple strains of bacteria growing to protect one another. But when Kevin Foster of the University of Oxford, UK, and his colleagues mixed two different strains of the opportunistic pathogen *Pseudomonas aeruginosa*, they found that the strains competed with each other, with one producing antibiotics called pyocins to kill its opponent. This mixture produced greater amounts of biofilm than did the individual strains, and introducing other antibiotics into the mix further stimulated biofilm formation.

The results suggest that the clinical use of antibiotics could be causing bacteria to make more biofilms, making the microbes harder to stamp out.

PLoS Biol. 13, **e1002191 (2015)**

SOCIAL SELECTION *Popular topics on social media*

Science reporting flaws exposed

Researchers often complain about inaccurate science stories in the popular press, but few air their grievances in a journal. Samuel Mehr, a PhD student at Harvard University in Cambridge, Massachusetts, discussed in a *Frontiers in Psychology* article some examples of media missteps from his own field — the effects of music on cognition. The opinion piece gained widespread attention online. "A very nice review of how media communicates science badly," tweeted Arseny Khakhalin, a neuroscientist at Bard College in Annandale-on-Hudson, New York. Seeing room for blame on both sides, Terry Hébert, a pharmacologist at McGill University in Montreal, Canada, posted on Facebook: "Journalists and scientists should both strive for accuracy in their claims."

Front. Psychol. 6, **00988 (2015)**

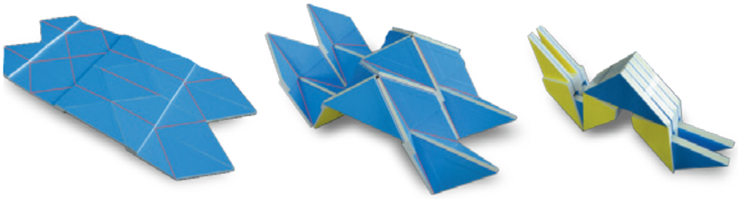
➲ NATURE.COM
For more on popular papers:
go.nature.com/awqdox

ENGINEERING

Origami for thick materials

Origami patterns designed for thin pieces of paper can be extended to thicker materials as well.

Previous attempts to fold 3D materials required adding layers of material or changing their geometry. To avoid this, Zhong You at the University of Oxford, UK, and his colleagues developed a method of assembling thick materials so that the hinges where they meet move in a limited number of ways. The researchers showed how carefully choosing the placement of the hinges and creases allows the structures to move and fold (**pictured**) in identical ways to origami patterns that use 2D materials.


SCIENCE/AAAS

The method could eventually improve the construction of foldable structures such as solar panels or aircraft wings, the authors report.

Science 349, **396–400 (2015)**

METABOLISM

Reroute bile for better metabolism

Altering the flow of bile in the mouse gut achieves the same weight-loss benefits as gastric-bypass surgeries.

In a gastric bypass, the stomach pouch is made smaller and connected to the middle of the small intestine. Naji Abumrad at Vanderbilt University Medical Center in Nashville, Tennessee, and his team tested whether the success of such procedures can be replicated by redirecting bile fluids, which break down dietary fat in the upper gut to enable fat absorption. Rerouting the bile duct to the ileum (the lower part of the small intestine) in obese mice resulted in weight loss and other metabolic improvements that were similar to those seen after a gastric bypass. The bile diversion reduced fat absorption and shifted the composition of gut microbes to be similar to that of lean mice.

Redirecting the bile duct in humans would be simpler than gastric bypass, but long-term safety studies are needed, the researchers say.

Nature Commun. 6, **7715 (2015)**

➲ NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

RESEARCH

Alzheimer's drugs

Alzheimer's patients enrolled in clinical trials of two antibody drugs showed minor improvements in their symptoms, two US drug companies reported on 22 July at the Alzheimer's Association International Conference in Washington DC. Eli Lilly's drug solanezumab slowed the rate of cognitive decline by 30% over two years. The drugs target amyloid-β protein in the brain, which accumulates in people with Alzheimer's. A small trial of Biogen's similar antibody, aducanumab, showed that high doses of the drug, but not moderate doses, removed amyloid and slowed symptoms. Both drugs have entered phase III trials. See page 509 for more.

Nearly Earth 2.0

NASA's Kepler spacecraft has found an extrasolar planet that is the closest thing yet to a true Earth analogue, the agency announced on 23 July. The planet, named Kepler-452b, orbits a bright star 430 parsecs (1,402 light years) away from Earth, within its star's habitable zone, where temperatures are right for liquid water to exist on the planet's surface. Kepler-452b is 60% larger than Earth and may be rocky. If it is, it will be the first terrestrial planet discovered within the habitable zone of a Sun-like star. (Similar planets have been found that orbit stars cooler and dimmer than the Sun.) See page 511 for more.

NASA/JHUAPL/SWRI

A hazy halo around Pluto

The New Horizons spacecraft has discovered hazes in Pluto's frigid atmosphere, NASA announced on 24 July. Fresh discoveries from the craft's historic fly-by on 14 July include possibly the first evidence that Pluto's atmosphere is beginning to freeze and fall as snow on the surface, which is what scientists had expected as its elliptical orbit takes it farther from the Sun. Powerful radio beams sent from Earth to the craft allowed researchers to measure the surprisingly low surface pressure of Pluto's atmosphere as it bent the radio waves. Other findings include glaciers of nitrogen ice flowing from the Sputnik Planum plains. See go.nature.com/bmmldk for more.

BUSINESS

Lifespan lineage

A genetic-ancestry company that boasts more than one million paying customers has teamed up with a Google-backed biotechnology firm to uncover lifespan-lengthening DNA. On 21 July, the genealogy firm Ancestry.com in Provo, Utah, announced a partnership between its subsidiary AncestryDNA and Calico, a research and development company in South San Francisco, California. Calico will use anonymized customer-genome data and ancestry records to search for genetic factors that contribute to longevity.

Generic drug deal

Jerusalem-based drug company Teva announced on 27 July that it would buy the generics arm of a rival company for around US$40.5 billion. Teva will pay $33.75 billion in cash and hand over roughly $6.75 billion in shares to Allergan, of Dublin, to acquire Allergan Generics. The companies say that they expect the deal to be completed early in 2016.

Hot drugs approved

Regulators have approved two members of a hotly pursued class of cholesterol drugs called PCSK9 inhibitors. On 21 July, Biotechnology firm Amgen of Thousand Oaks, California, announced that the European Commission had approved its drug, called evolocumab. Three days later, the US Food and Drug Administration approved Praluent (alirocumab), made by Sanofi US of Bridgewater, New Jersey, and Regeneron of Tarrytown, New York. Both drugs are approved for use in people with familial hypercholesterolaemia, a genetic condition that causes high cholesterol (see *Nature* **496,** 152–155; 2013).

EVENTS

Name that asteroid

The Japan Aerospace Exploration Agency (JAXA) has invited the public to suggest names for the asteroid 1999 JU3, which JAXA's probe Hayabusa-2 will reach in 2018. The spacecraft, which launched in December 2014, will collect samples from the rock and return to Earth in 2020 (see *Nature* http://doi.org/6dg; 2014). Candidate names will be vetted by a panel of specialists, the asteroid's discoverers and, ultimately, the International Astronomical Union. The contest, which opened on 22 July, resembles NASA's call to name features on Pluto; submissions close on 31 August (see go.nature.com/rybeqw).

Outbreak ending

The South Korean outbreak of Middle East respiratory syndrome (MERS) is set to be declared as over. Beginning in hospitals in mid-May, the coronavirus outbreak was quickly brought under control, with case numbers peaking on 1 June. In total, 186 people were confirmed as infected, 36 of whom have died. No new cases have occurred since 2 July. On 27 July, the last of 16,693 contacts was released from isolation. The outbreak can officially be declared over when no new cases have been detected for 28 days.

Weapons warning

Nearly 2,000 researchers and commentators released an open letter on 28 July calling for the United Nations to ban autonomous weapons. The letter warns that weapons using artificial intelligence (AI) may be only years away, and that their creation could lead to an "AI arms race". It was released at the International Joint Conference on Artificial Intelligence in Buenos Aires, Argentina, and signatories include more than 1,000 AI researchers, as well as physicist Stephen Hawking and inventor Elon Musk.

Nigeria polio-free

On 24 July, Nigeria marked one year since its last case of polio caused by wild virus. The West African nation will be removed from the list of polio-endemic countries once laboratory tests from across the country are complete. Pakistan and Afghanistan are now the only two countries where transmission of polio virus has never been halted. Africa could be certified polio-free as soon as August 2017, three years after the continent's last case, in Somalia, which was caused by a virus that spread from Nigeria. See go.nature.com/uomxdr for more.

JOHN MACDOUGALL/AFP/GETTY

PEOPLE

Pachauri out

Rajendra Pachauri, the former chairman of the Intergovernmental Panel on Climate Change (IPCC), has been replaced as director-general of The Energy and Resources Institute (TERI) in New Delhi, India. The action by TERI's governing council, announced on 23 July, comes on the heels of sexual-harassment allegations that cost Pachauri (**pictured**) his post at the IPCC in February. He will be replaced by Ajay Mathur, an energy official in the Indian government. In a press release, TERI stated that the search for Pachauri's replacement began in September 2014.

Genome head quits

The revered leader of a leading genomics research centre in China has abruptly stepped down. Jun Wang is leaving his post as chief executive of BGI in Shenzhen, the institute announced on 17 July, and will pursue research into artificial intelligence. Wang has worked at BGI since it opened in 1999. He is credited with leading some of its major accomplishments, including being the first to sequence the genome of an Asian person, the giant panda and the human gut microbiota. He also led contributions to the Human Genome Project and the initiative to sequence the rice genome. See go.nature.com/nhdkjs for more.

POLICY

Pesticide ban lifted

A UK government agency has used emergency rules to make controversial neonicotinoid insecticides available to some farmers, despite a European ban imposed on the chemicals in 2013. Neonicotinoids have been linked to declines in bee populations in numerous scientific studies. The UK Department for Environment, Food & Rural Affairs authorized limited emergency use of two pesticides on 22 July after a request by farmers. The National Farmers' Union says that the pesticides are needed to protect around 300 square kilometres of oilseed rape in England from cabbage stem flea beetles. See go.nature.com/ytp9q6 for more.

COMING UP

2–6 AUGUST
The 27th International Congress for Conservation Biology and the 4th European Congress for Conservation Biology, join in Montpellier, France, to discuss ways to tackle conservation.
go.nature.com/rsu6x3

4–6 AUGUST
Planetary scientists gather in Arcadia, California, to argue the case for their favourite Martian landing site. Around 30 candidate locations are in the running to be chosen for NASA's 2020 Mars rover.
go.nature.com/x4wh7c

3–14 AUGUST
Honolulu, Hawaii, is awash with 3,500 astronomers, when the International Astronomical Union holds its general assembly.
go.nature.com/4uj4ia

Malaria vaccine

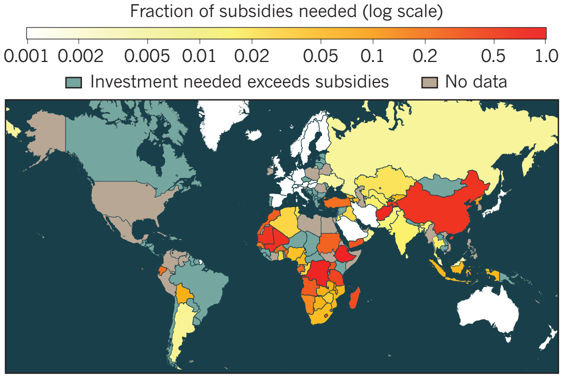
The world's first vaccine against malaria can now be used routinely in children — but only in countries where the disease is endemic. On 24 July, the European Medicines Agency in London declared Mosquirix safe for use and moderately effective against the malaria parasite *Plasmodium falciparum*, in combination with established protective measures such as bed nets. The World Health Organization must now formally agree to recommend its use in children. Mosquirix has been developed by UK-based drug firm GlaxoSmithKline, and is also protective against hepatitis B.

TREND WATCH

Ending fossil-fuel subsidies, which cost US$550 billion globally in 2011, could free up money to achieve universal access to basic services such as clean water, according to researchers at the Mercator Institute on Global Commons and Climate Change in Berlin (M. Jakob *et al. Nature Clim. Change* **5,** 709–712; 2015). The team says that access to clean water could be achieved for $190 billion, with sanitation and electricity costing $370 billion and $430 billion, respectively.

SHARING SUBSIDIES

The portion of national fossil-fuel subsidies that would need to be redistributed to provide clean water by 2030 is highest in China and in parts of Africa.


Fraction of subsidies needed (log scale)
0.001 0.002 0.005 0.01 0.02 0.05 0.1 0.2 0.5 1.0
■ Investment needed exceeds subsidies
■ No data

SOURCE: M. JAKOB *ET AL. NATURE CLIM. CHANGE* **5**, 709–712 (2015)

➲ NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

RAUL ARBOLEDA/AFP/GETTY

Alzheimer's disease is marked by cognitive decline and the accumulation of proteins in the brain.

DRUG DEVELOPMENT

Alzheimer's drugs show progress

Protein-targeting antibodies succeed after many failures.

BY SARA REARDON

For years, scientists studying Alzheimer's disease have been frustrated on two fronts. They have struggled to understand whether the amyloid-β protein that accumulates in sufferers' brains is a driver of the disease, or just a symptom. And without a clear understanding of the condition's cause, they have searched fruitlessly for effective treatments.

The latest results from clinical trials of two antibody drugs could provide a path forward. For the first time, such drugs — which target amyloid — seem to have slowed the progression of the disease. The findings, released on 22 July at the Alzheimer's Association International Conference, support the idea that amyloid deposits cause the mental deterioration seen in people with Alzheimer's.

"We're creeping in the right direction," says Samuel Gandy, a neurobiologist at Mount Sinai School of Medicine in New York. "A lot of the euphoria is because things were so negative for so long." Still, many researchers doubt whether the minor improvements reported last week will hold up in larger trials.

Pharmaceutical firm Eli Lilly of Indianapolis, Indiana, says that in a trial with 440 participants, its drug solanezumab seemed to slow the cognitive decline of people with mild Alzheimer's by about 30%. Over 18 months, people in the treatment group experienced a loss of mental acuity equivalent to the deterioration experienced in just 12 months by those in a placebo group with Alzheimer's disease of similar severity.

Lilly snatched this small victory from the jaws of defeat. In 2012, the company reported no difference between patients who had taken solanezumab for 18 months and those who had received a placebo. But when the company reanalysed the trial data, it found a slight improvement in participants whose symptoms were mild when the trial began. Lilly continued the test for six months and began giving solanezumab to the 440-member control group, whose disease was by then more advanced.

The latest results show that cognitive decline in the 'late start' group slowed to match the rate seen in the 440 people who had been treated for the entire study. This suggests that solanezumab targeted the root of Alzheimer's disease.

Drug-maker Biogen, of Washington DC, presented results that show a moderate dose of its drug aducanumab reduced amyloid build-up in 23 people, but did not have statistically significant clinical benefits. In March, the company reported that 27 people who received high doses of aducanumab for one year showed significantly less cognitive decline than people who received a placebo, and had less amyloid in their brains.

Many experts are greeting these results with tempered excitement, given the relatively small size of the clinical trials. But Eric Siemers, an Alzheimer's researcher at Lilly, is more optimistic. "It's surprising to me that [solanezumab] worked so well," he says. "There's a lot of promise to slow progression."

Lilly launched a larger phase III trial of solanezumab in 2013, enrolling 2,100 people with mild symptoms and amyloid deposits in their brains. The study will end in October 2016. And last December, Biogen said that it would launch a phase III trial with 2,700 participants that would run for 18 months. ▶

Lon Schneider, an Alzheimer's researcher at the University of Southern California in Los Angeles, questions the decision to start large trials before the drugs, and the amyloid hypothesis, have been well validated. "Why are there so many antibodies when none so far have proven efficacy?" he asks, noting that behavioural interventions, such as diet and exercise, have been shown to slow Alzheimer's as much as any drug (A. M. Clarfeld and T. Dwolatzky *JAMA Intern. Med.* **173,** 901–902; 2013).

But not everyone agrees. "This is the time to be bold," says Randall Bateman, a neurologist at Washington University in St. Louis, Missouri. "It seems to me the cost of delay from a human-suffering standpoint is much more expensive than the cost of moving forward."

Bateman is leading a trial that is testing solanezumab and gantenerumab — developed by Roche of Basel, Switzerland — in 160 people between 18 and 80 years old who have a genetic risk of Alzheimer's, but no symptoms. It is one of several efforts attempting to determine whether the disease can be prevented by destroying amyloid protein before the brain is damaged. That harm occurs over decades (R. J. Bateman *et al. N. Engl. J. Med.* **367,** 895–804; 2012), and many Alzheimer's researchers suspect that antibody-drug trials have failed because they have treated people too late.

This hypothesis is supported by Lilly's finding that only people with mild disease benefit from solanezumab. The latest results also demonstrate for the first time in humans that slowing amyloid deposition can slow down cognitive decline, says Eric Reiman, executive director of the Banner Alzheimer's Institute in Phoenix, Arizona.

That is important because the US Food and Drug Administration has said that it will not approve drugs that block amyloid deposits without sufficient evidence of a clinical benefit. If one drug company can prove the cause and effect between amyloid accumulation and Alzheimer's progression, all companies will benefit, says Reiman, who is leading a trial of crenezumab, a Roche drug that has also failed previously in large trials.

If such drugs falter in larger preventative trials, that would be a setback for Alzheimer's research in general, says Gandy. "The main concern is that the pipeline behind amyloid-reducing agents is really pretty spare," he says. However, at least three companies are developing treatments — some of which are antibody drugs — that target a different protein, tau, which destroys neurons in advanced Alzheimer's disease. ■

See go.nature.com/lathpu for a longer version of this story.

The MV *Cape Race* is using sonar to map the depth of water around Greenland's west coast.

MARIA STENZEL/UCI

GLACIOLOGY

NASA launches mission to Greenland

Ship and planes will probe water–ice interface in fjords.

BY JEFF TOLLEFSON

When the retired fishing trawler MV *Cape Race* sets off along Greenland's west coast this week, it will start hauling in a scientific catch that promises to improve projections of how the ice-covered island will fare in a warming world. The ship's cruise is the initial phase of a six-year air and sea campaign to probe interactions between Greenland's glaciers and the deep, narrow fjords where they come to an end.

Called Oceans Melting Greenland (OMG), the US$30-million NASA project will help scientists to predict the future of the Greenland ice sheet, which holds enough water to boost sea levels by around 6 metres and already seems to be melting more rapidly in response to increasing air temperatures. But it is not clear how much the oceans affect the rate of melting along the island's edges, which depends on poorly known variables such as how warm, saline water interacts with the glaciers.

"It should be a powerful constraint on our knowledge and ability to model ice loss there," says principal investigator Joshua Willis, an oceanographer at NASA's Jet Propulsion Laboratory in Pasadena, California.

When simulating glacier dynamics, current global climate models consider only ice's interactions with the atmosphere, says William Lipscomb, an ice modeller at Los Alamos National Laboratory in New Mexico. He is working to incorporate ice–ocean interactions around Antarctica into a climate model being developed by the US Department of Energy. But in Greenland, the intricately carved coastline makes this much more difficult. The department plans to give researchers at the Naval Postgraduate School in Monterey, California, $466,000 over 2 years to build a detailed model that will link the land ice and oceans around Greenland. OMG data will help to validate that model, says project leader Frank Giraldo.

Work by OMG participant Eric Rignot, a glaciologist at the University of California, Irvine, underscores the importance of detailed data (E. Rignot *et al. Geophys. Res. Lett.* http://doi.org/6dn; 2015). Using sonar data from one part of western Greenland, Rignot's team found that existing maps underestimate the depth of three fjords by several hundred metres. It also found that glaciers flowing into all three fjords extended deeper than was thought, far enough below fresh surface waters to reach a warm, salty layer flowing up from the Atlantic Ocean that could accelerate melting and contribute more to sea-level rise than had been believed.

"With OMG, we are going to reveal the depth of these fjords," says Rignot.

The programme will also provide valuable information on the physical characteristics of glacier ice. Last December, geophysicist Beata Csatho of the University at Buffalo in New York and her colleagues reported using surface-elevation data to estimate how much ice mass Greenland had lost between 1993 and 2012 (B. M. Csatho *et al. Proc. Natl Acad. Sci. USA* **111,** 18478–18483; 2014). The data were fairly reliable over the island's interior, Csatho says, but measurements were more difficult along its edges, where the ice tends to be warmer, thicker and full of crevices. "It's still a challenge to get the mass of these glaciers," she says.

When the aerial phase of OMG begins next year, planes will fly inland from the coast, taking measurements of slight changes in gravitational pull that can be used to produce low-resolution maps of the topography under both water and ice. Planes will also drop more than 200 temperature and salinity probes into fjords and coastal waters, and take radar measurements along the coast to track large-scale ice loss over five years. Analysing that ice loss in light of the new topographical and oceanographic data will help researchers to determine where, and to what extent, deeper saltwater currents affect glaciers.

Lipscomb says that all these OMG data should help modellers as they incorporate ocean–ice interactions around Greenland into their models. That work is still in its early stages, he says, "but the data that they are getting in this project is exactly what we need". ■

PLANETARY SCIENCE

Kepler spies most Earth-like planet yet

NASA mission finds a potentially rocky world orbiting a star that resembles the Sun.

BY ALEXANDRA WITZE

Even as astronomers are reporting what looks like the closest thing yet to an Earth-like exoplanet, NASA is winding down the prolific Kepler mission that made the find. Sometime next year, team scientists plan to release their final list of planet discoveries.

On 23 July, the Kepler team announced the existence of a planet 1.6 times the size of Earth, orbiting a Sun-like star 430 parsecs away (J. M. Jenkins *et al. Astron. J.* **150,** 56, 2015). The planet, named Kepler-452b, is in the habitable zone, orbiting its star at a distance where liquid water could exist. Team scientists say that there is a little more than a 50% chance that the planet is rocky, which would make it the closest thing to a true Earth analogue yet discovered.

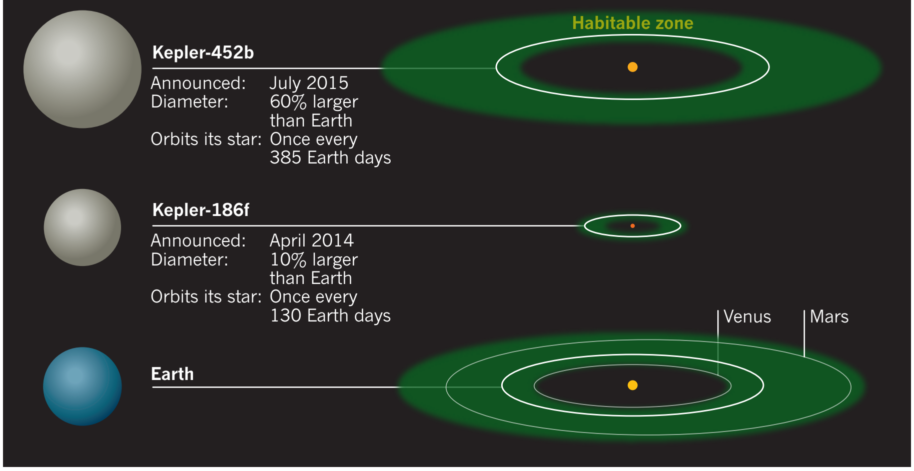
Kepler's latest batch of discoveries also includes at least 11 other planets that are all less than twice the size of Earth and orbit in their stars' habitable zones. But Kepler-452b's star is slightly brighter than the Sun, in contrast to the cool, dim stars that host other known Earth-sized planets.

"It is the first terrestrial planet in the habitable zone around a star very similar to the Sun," says Douglas Caldwell, an astronomer at the SETI Institute in Mountain View, California.

Scientists cannot measure the mass of Kepler-452b directly, but modelling suggests that the planet is probably five times the mass of Earth. It whirls around its star once every 385 days, tantalizingly close to Earth's 365-day year (see 'Habitable hunt').

HABITABLE HUNT

NASA's Kepler mission has found several small planets outside the Solar System that could harbour life. They orbit their stars in the 'habitable zone' where liquid water could exist.


Habitable zone
Kepler-452b
Announced: July 2015
Diameter: 60% larger than Earth
Orbits its star: Once every 385 Earth days
Kepler-186f
Announced: April 2014
Diameter: 10% larger than Earth
Orbits its star: Once every 130 Earth days
Venus
Mars
Earth

SOURCE: NASA

The planet's star is about 1.5 billion years older than the 4.5-billion-year-old Sun, and Kepler-452b is about the same age. During its first 5 billion years, the planet would have received less energy from its star than Earth does from the Sun, but it may now offer a glimpse of Earth's future. Kepler-452b's star is growing hotter and brighter as part of its natural evolution, so anyone living on the planet would see their world drying out — just as Earth will as the Sun evolves.

From 2009 to 2013, the Kepler craft stared at a small patch of sky looking for slight decreases in starlight that signalled a planet moving, or transiting, across the face of its star. It stopped taking data when it was crippled by a broken reaction wheel that guided its pointing, but engineers later revived it in a limited fashion.

The craft has discovered more than 1,000 confirmed planets, including Kepler-452b, and more than 4,660 candidates.

In the next year, Kepler scientists will perform analyses to reduce the signal-to-noise ratio in their data, teasing out as many planets as possible. "Imagine you're trotting through the weeds in a field, looking for precious stones on the ground," says Natalie Batalha, Kepler's mission scientist and an astronomer at NASA's Ames Research Center in Moffett Field, California. "We're going through with a lawnmower so that the stones are easier to see."

In January, the European Southern Observatory began a search for transiting planets from its telescopes in Chile. NASA plans to launch a space-based successor to Kepler called the Transiting Exoplanet Survey Satellite in 2017. ■

ANTHROPOLOGY

Neanderthals had outsize effect on human biology

From skin disorders to the immune system, sex with archaic species changed Homo sapiens.

BY EWEN CALLAWAY

Our ancestors were not a picky bunch. Overwhelming genetic evidence shows that *Homo sapiens* had sex with Neanderthals, Denisovans and other archaic relatives. Now researchers are using large genomics studies to unravel the decidedly mixed contributions that these ancient romps made to human biology — from the ability of *H. sapiens* to cope with environments outside Africa, to the tendency of modern humans to get asthma, skin diseases and maybe even depression.

The proportion of the human genome that comes from archaic relatives is small. The genomes of most Europeans and Asians are 2–4% Neanderthal[1], with Denisovan DNA making up about 5% of the genomes of Melanesians[2] and Aboriginal Australians[3]. DNA slivers from other distant relatives probably pepper a variety of human genomes[4].

But these sequences may have had an outsize effect on human biology. In some cases, they are very different from the corresponding *H. sapiens* DNA, notes population geneticist David Reich of Harvard Medical School in Boston, Massachusetts — which makes it more likely that they could introduce useful traits. "Even though it's only a couple or a few per cent of ancestry, that ancestry was sufficiently distant that it punched above its weight," he says.

Last year, Reich co-led one of two teams that catalogued the Neanderthal DNA living on in modern-day humans[5,6]. The studies hinted that Neanderthal versions of some genes may have helped Eurasians to reduce heat loss or grow thicker hair. But the evidence that these genes were beneficial was fairly weak.

XIN LU/GETTY

A gene variant from archaic humans helps modern-day Tibetans to cope with high altitudes.

To get a better handle on how Neanderthal DNA shapes human biology, Corinne Simonti and Tony Capra, evolutionary geneticists at Vanderbilt University in Nashville, Tennessee, turned to genome-wide association studies (GWAS) that had already compared thousands of DNA variants in people with and without a certain disease or condition.

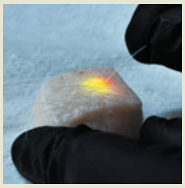
Using de-identified genome data and medical records from 28,000 hospital patients, Simonti and Capra looked for differences in traits and medical diagnoses between people with a particular Neanderthal gene variant and those with the *H. sapiens* version of the same gene. They found that the Neanderthal variants seemed to slightly increase the risk of conditions such as osteoporosis, blood-coagulation disorders and nicotine addiction. Another analysis, which looked at the combined effects of many DNA variants, painted a more mixed picture. It revealed links between Neanderthal DNA and depression, obesity and certain skin disorders, with some variants being associated with an increased risk and others with a reduced risk. Simonti presented the data at the annual meeting of the Society for Molecular Biology and Evolution in Vienna on 15 July.

Neanderthal gene variants, like most human variants, had only a tiny effect on the risk of developing these conditions, notes Capra. But seeing Neanderthal genes involved in skin disorders, including lesions triggered by sun exposure, chimes with previous studies linking Neanderthal DNA to skin biology, he says.

In some cases, the effects of the archaic genes may have changed over the ages. Simonti and Capra also reported at the Vienna meeting that blood-coagulation disorders experienced by modern humans could be related to Neanderthal immune genes, although previous studies

MORE ONLINE

TOP STORY

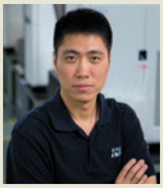

MATJAŽ HUMAR/SEOK HYUN YUN

Droplets of oil turn cells into tiny lasers go.nature.com/4vf5fa

MORE NEWS

- Brain area found that may make humans unique go.nature.com/xxn9p2
- Simple cooking methods flush arsenic out of rice go.nature.com/vl6367
- Scientists unite to warn against flawed chemical reagents go.nature.com/dpttdu

Q&A


BGI

Genomics pioneer Jun Wang on his new venture in artificial intelligence go.nature.com/pupjhy

have suggested that archaic immune genes may have helped *H. sapiens* to cope with diseases that they encountered outside Africa.

Also at the meeting, a team led by Michael Dannemann, a computational biologist at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany, reported that many humans have Neanderthal and Denisovan versions of genes that encode proteins called toll-like receptors (TLRs), which sense pathogens and launch a rapid immune response. Furthermore, cultured human cells containing the archaic versions tended to express the TLRs at higher levels than cells with the *H. sapiens* versions[7]. Although previous GWAS linked the archaic versions to a reduced risk of *Heliobacter pylori* infection, which can cause stomach ulcers, the variants were also associated with higher rates of allergies.

"Many traits that were adaptive 10,000 years ago might be maladaptive today" because of lifestyle, diet and other shifts, notes Rasmus Nielsen, a population geneticist at the University of California, Berkeley.

At least one archaic trait has clear benefits in contemporary humans. Last year, Nielsen's team reported that the Denisovan-like version of a gene called *EPAS1* helps modern Tibetans to cope with life at altitudes of 4,000 metres, by preventing their blood from thickening[8].

Many researchers see Nielsen's *EPAS1* discovery as the poster child for humans' archaic biology, because the benefits of the Denisovan version are so clear-cut. But proving such insights rests on laborious studies, including engineering mice to carry the archaic mutations and exhaustively testing the animals' biology, notes Reich. "Each new finding is going to be very hard won." ■

1. Green, R. E. *et al. Science* **328,** 710–722 (2010).
2. Meyer, M. *et al. Science* **338,** 222–226 (2012).
3. Rasmuseen, M. *et al. Science* **334,** 94–98 (2011).
4. Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C. & Wall, J. D. *Proc. Natl Acad. Sci. USA* **108,** 15123–15128 (2011).
5. Vernot, B. & Akey, J. M. *Science* **343,** 1017–1021 (2014).
6. Sankararaman, S. *et al. Nature* **507,** 354–357 (2014).
7. Dannemann, M., Andrés, A. M. & Kelso, J. Preprint at http://dx.doi.org/10.1101/022699 (2015).
8. Huerta-Sánchez, E. *et al. Nature* **512,** 194–197 (2014).

POLITICS

Budget showdown leaves US science agencies in limbo

Lawmakers face looming deadline to reach a deal — or risk government shutdown.

BY CHRIS CESARE

When the US Congress returns from its late-summer recess in early September, lawmakers and President Barack Obama will have less than three weeks to reach a budget deal for 2016, and in doing so determine the funding of key science agencies.

The most likely scenario, experts say, is that a temporary deal will be made to keep the government operating for weeks or months after the 2016 fiscal year begins on 1 October. That is cold comfort for US science agencies and researchers who have endured years of bruising partisan spending battles.

"We're basically headed into a period of frustration where nothing's going to happen for a couple months, and we're just going to have to deal with it," says Jennifer Zeitzer, director of legislative relations at the Federation of American Societies for Experimental Biology in Bethesda, Maryland.

The current funding agreement expires on 30 September. And, in protest against legislators' embrace of the across-the-board budget cuts known as sequestration, Obama has threatened to veto many of the 2016 spending bills introduced by the Republican-controlled House of Representatives and Senate.

Some agencies, such as the National Institutes of Health (NIH), seem likely to emerge as winners in any deal (see 'Budget battle'). The House has proposed increasing the NIH budget by US$1.1 billion in 2016, to $31.2 billion; the Senate's proposal of $32.1 billion is even more generous. And both are close to the White House proposal of $31.3 billion.

But for other agencies and programmes, the prospects are not so clear. The House matched Obama's $18.5-billion request for NASA, and the Senate is close at $18.3 billion. But the House wants to chop $101 million from the space agency's $1.8-billion Earth-sciences account, which funds research on topics such as climate change. The Senate would boost the wing's funding by roughly $148 million. Similarly, the House bill would set aside $5.2 billion for the National Oceanic and Atmospheric Administration ▶

BUDGET BATTLE

How funding proposals from the White House and Congress stack up, by agency (US$ millions).

Agency	2015 estimated budget	2016 White House request	2016 House bill	2016 Senate bill
Biomedical research and public health				
National Institutes of Health	30,311	31,311	31,184	32,084
Centers for Disease Control and Prevention	6,073	6,170	7,010	6,711
Food and Drug Administration	2,596	2,744	2,619	2,629
Physical sciences				
National Science Foundation	7,344	7,724	7,394	7,344
NASA (science)	5,245	5,289	5,238	5,295
Department of Energy Office of Science	5,068	5,340	5,100	5,100
National Institute of Standards and Technology	864	1,120	855	893
Earth and environment				
Environmental Protection Agency	8,140	8,592	7,422	7,597
National Oceanic and Atmospheric Administration	5,449	5,983	5,167	5,382
US Geological Survey	1,045	1,195	1,045	1,059

SOURCES: WHITE HOUSE, US CONGRESS

▶ (NOAA), 5% less than the current level and about $800 million short of Obama's request. The Senate bill would reduce NOAA spending by just 1%.

But it is the National Science Foundation (NSF) that has most polarized lawmakers. The House's NSF spending bill would require the agency to award 70% of its $6-billion research fund to biology, computer science, engineering, mathematics and the physical sciences. The unusual provision would effectively impose a 16% cut to geoscience and social-sciences programmes, according to an analysis by the American Institute of Physics. By contrast, the Senate's bill does not set funding levels for particular disciplines.

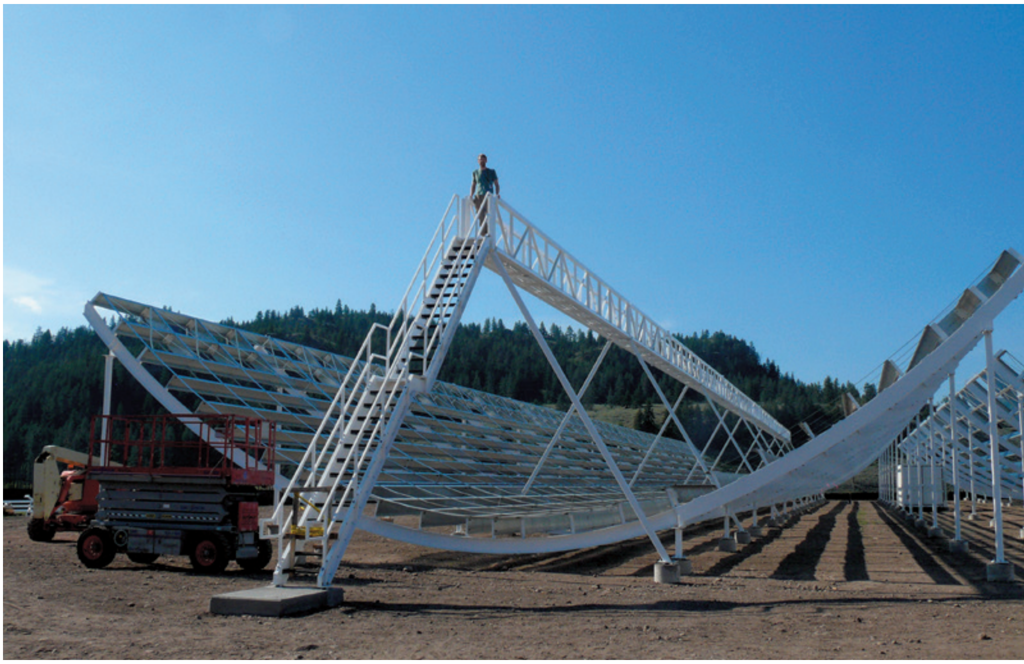
BASIC FOCUS

Powerful House Republicans, most notably science-committee chair Lamar Smith of Texas, have argued that the NSF should concentrate on basic research. Smith has also tried to highlight what he sees as questionable grants by the science agency, such as funding for a study of mental health in Nepal. But Gloria Waters, vice-president and associate provost for research at Boston University in Massachusetts, says that legislators often misunderstand the role of basic science. "People have this idea that science funding should go to something that should have an immediate and direct impact on society, but that's not how science works," she says.

Deciding which projects to fund is made more difficult by a lack of money, says Hannah Carey, a physiologist at the University of Wisconsin–Madison. "I've experienced it — you put in a grant to continue your work that gets a very, very good score and would have been funded in a better climate," says Carey, who spent a year working as a programme director in the NSF's biosciences division. "It's disheartening."

A short-term spending deal would avert a government shutdown of the sort that ground most research to a halt in October 2013. But a stopgap arrangement could still make life difficult for researchers. Such deals generally prevent agencies from starting new programmes or ending old ones without specific authorization from Congress. And agencies can face unexpected budget shortfalls if lawmakers agree to cut spending after months of operating under a temporary funding agreement, because the cuts would be retroactive.

For now, scientists are left to wait while the negotiations between Congress and the White House play out. Zeitzer says that a deal may not be struck until the last minute. "I'm hearing there's a real good chance they'll take us to the brink," she says. ■



The CHIME telescope array will search for a particular kind of hydrogen emission from ancient galaxies.

MARK HALPERN/CHIME COLLABORATION

COSMOLOGY

Half-pipe array to map teen Universe

Canadian telescope aims to chart cosmic expansion rate between 10 billion and 8 billion years ago.

BY DAVIDE CASTELVECCHI

It sounds almost too apt to be true. An observatory shaped like the half-pipes used by snowboarders, and dependent on technology originally designed for gaming and mobile phones, will soon be tasked with plugging a crucial gap in the cosmological record: what the Universe did when it was a teenager.

The information will allow cosmologists to gauge whether the strength of dark energy — the force accelerating the Universe's expansion — has changed over time, an unresolved question that governs the fate of the cosmos.

Whereas typical radio telescopes have round dishes, the Canadian Hydrogen Intensity Mapping Experiment (CHIME) comprises four 100-metre-long, semi-cylindrical antennas, which lie near the town of Penticton in British Columbia.

From 2016, CHIME's half-pipes, which are scheduled to be completed this week, will detect radio waves emitted by hydrogen in distant galaxies. These observations would be the first measurements of the Universe's expansion rate between 10 billion and 8 billion years ago, a period in which the cosmos went "from being a kid to an adult", says Mark Halpern, the leader of CHIME and an experimental cosmologist at the University of British Columbia in Vancouver. Straight after the Big Bang 13.8 billion years ago, the rate of the Universe's expansion slowed. But somewhere during the 'adolescent' period, dark energy — which eventually turned the Universe's slowing expansion into the acceleration observed today — began to be felt, he says.

It is a window in time that has, until now, been closed. Cosmologists measure the Universe's past expansion rate using ancient objects, such as supernova explosions and the voids between galaxies, that are so distant that their light is only now reaching Earth. Over the past few decades, such objects have revealed that the cosmos has been expanding at an accelerating rate for more than 6 billion years. And surveys of quasars — mysterious, super-bright objects that outshine the entire galaxies they lie in — have shown that until 10 billion years or so ago, the Universe's expansion was slowing down.

But cosmologists have struggled to measure the expansion rate in the interim, leaving open the question of whether the strength of dark energy's repulsive force may have varied over time.

CHIME is designed to fill the gap, says Kendrick Smith, an astrophysicist at the Perimeter Institute for Theoretical Physics in Waterloo, Canada, who will work on analysing CHIME's data. The half-pipe antennas will allow CHIME to receive radio waves coming from anywhere along a narrow, straight region of the sky at any given time. "As the Earth rotates, this straight shape sweeps out the sky," says Smith.

To sort out where individual signals are coming from, a custom-built supercomputer made of 1,000 relatively cheap graphics-processing units — the type used for high-end computer gaming — will crunch through nearly 1 terabyte of data per second. The team will also use signal amplifiers originally developed for mobile phones. Without such powerful consumer-electronics components, CHIME would have been prohibitively expensive, says experimental cosmologist Keith Vanderlinde of the University of Toronto, Canada, who is co-leading the project.

CHIME's supercomputer will look specifically for radio waves with a wavelength that suggests an age of 11 billion to 7 billion years, emitted by the hydrogen in the interstellar space inside galaxies (at their source, such emissions have a wavelength of 21 centimetres). Researchers will then subtract the 'radio noise' in the same wavelength range that comes from the Milky Way and Earth.

Although CHIME will not be able to distinguish individual galaxies in this way, clumps of hundreds or thousands of galaxies will show up, says Vanderlinde. This will allow researchers to map the expansion rate of the voids between the clumps, and in turn to calculate the strength of dark energy during that time.

"As the Earth rotates, this straight shape sweeps out the sky."

If the results imply that the strength of dark energy then was the same as it has been in the past 6 billion years, it could suggest that galaxies will eventually lose sight of each other. But if the strength of dark energy has changed over the eons, all bets are off: the Universe could collapse in a 'big crunch', for example, or be ripped apart into its subatomic components.

As well as mapping the adolescent Universe, CHIME could also detect hundreds of the mysterious 'fast radio bursts' that last just milliseconds and have no known astrophysical explanation. And it will help other experiments to calibrate measurements of radio waves from rapidly spinning neutron stars, which researchers hope to use to detect the ripples in space-time known as gravitational waves (see *Nature* **463,** 147; 2010).

CHIME is part of a growing trend in astronomy. A number of experiments that are now active or in the planning stages, including the hotly anticipated Square Kilometer Array — to be built on sites in Australia and South Africa — are designed to look at hydrogen emissions with 21-centimetre wavelengths. These emissions are an untapped trove of cosmological information, says Tzu-Ching Chang, an astrophysicist at the Academia Sinica Institute of Astronomy and Astrophysics in Taipei who helped to pioneer the hydrogen mapping of galaxies in a 2010 study (T.-Z. Chang *et al. Nature* **466,** 463–465; 2010). She likens the boom in hydrogen mapping today to the trend in the 1990s of studying the relic radiation of the Big Bang, which revolutionized cosmology. ■

CLARIFICATION

The News story 'US tailored-medicine project aims for ethnic balance' (*Nature* **523,** 391–392; 2015) implied that the plan for the whole Precision Medicine Initiative is due to be announced in the next few weeks. Actually, the forthcoming plan is just for the cohort-study component of the project.

T. C. WORLEY/NYT/REDUX/EYEVINE

The trouble with CHECKLISTS

An easy method that promised to save lives in hospitals worldwide may not be so simple after all.

BY EMILY ANTHES

Many hospitals have introduced pre-surgery checklists, with mixed results.

Before making the first incision, confirm the patient's identity. Mark the surgical site. Ask about allergies. Discuss any anticipated blood loss. Introduce yourself by name. These are some of the 19 tasks on the World Health Organization (WHO) Surgical Safety Checklist, a simple list of actions to be completed before an operation in order to cut errors and save lives.

In 2007 and 2008, surgical staff at eight hospitals around the world tested the checklist in a pilot study[1]. The results were remarkable. Complications such as infections after surgery fell by more than one-third, and death rates dropped by almost half. The WHO recommended that all hospitals adopt its checklist or something similar, and many did.

The UK National Health Service (NHS) immediately required all of its treatment centres to put the checklist into daily practice; by 2012, nearly 2,000 institutions worldwide had tried it. The idea of checklists as a simple and cheap way to save lives has taken hold throughout the clinical community. It has some dynamic champions, including Atul Gawande, a surgeon at Brigham and Women's Hospital in Boston, Massachusetts, who led the pilot study and has spread the word through talks, magazine articles and a best-selling book, *The Checklist Manifesto* (Metropolitan, 2009).

But this success story is beginning to look more complicated: some hospitals have been unable to replicate the impressive results of initial trials. An analysis of more than 200,000 procedures at 101 hospitals in Ontario, Canada, for example, found no significant reductions in complications or deaths after surgical-safety checklists were introduced[2]. "We see this all the time," says David Urbach, a surgeon at the University of Toronto who led the Ontario analysis. "A lot of studies that should be a slam dunk don't seem to work in practice." The stakes are high, because poor use of checklists means that people may be dying unnecessarily.

A cadre of researchers is working to make sense of the discrepancies. They are finding a variety of factors that can influence a checklist's success or failure, ranging from the attitudes of staff to the ways that administrators introduce the tool. The research is part of the growing field of implementation science, which examines why some innovations that work wonderfully in experimental trials tend to fall flat in the real world. The results could help to improve the introduction of other evidence-based programmes, in medicine and beyond.

"We need to learn the lessons from programmes and interventions like the checklist so we don't make the same mistakes again," says Nick Sevdalis, an implementation scientist at King's College London.

REPLICATION FRUSTRATION

One of the first to demonstrate the potential of checklists in health care was Peter Pronovost, an anaesthesiologist and critical-care physician at Johns Hopkins University School of Medicine in Baltimore, Maryland. In 2001, Pronovost introduced a short checklist for health-care workers who insert central venous catheters, or central lines, which are often used in an intensive care unit (ICU) to test blood or administer drugs. The trial showed that asking practitioners to confirm that they had performed certain simple actions, such as washing their hands and sterilizing the insertion site, contributed to a dramatic reduction in the risk of life-threatening infections[3]. The list got a larger test in a now-famous trial[4] known as the Keystone ICU project, launched in Michigan in October 2003. Within 18 months, the rate of catheter-related bloodstream infections fell by 66%.

Checklists were not completely new to medicine, but Pronovost's work attracted attention because it suggested that they could save lives. Gawande penned an inspiring feature in *The New Yorker*[5], asking: "If something so simple can transform intensive care, what else can it do?" Checklists began to proliferate. Now there are checklists for procedures involving anaesthesia, mechanical ventilation, childbirth and swine flu. Many studies have generated promising results, showing that the lists improve patient outcomes in hospitals from Norway to Iran.

But there have also been some failures. This January, less than a year after the report from Ontario, a different team of scientists reported[6] that a surgical checklist modelled on Pronovost's list did not improve outcomes at Michigan hospitals. And although the central-line checklist for ICUs has provided lasting benefits in Michigan, a British initiative called Matching Michigan, which aimed to replicate the Keystone programme, seemed to make no difference to infection rates[7].

Some experts suspect that the failure to replicate could be a matter of how the initial trials or the follow-up studies were designed. Gawande's pilot study of the WHO surgical checklist, for example, was not randomized and had no control group. Instead, it compared complication and death rates before and after the checklist was introduced. Critics say that this makes it difficult to determine what other factors might have influenced outcomes.

Gawande acknowledges the limitation, which was due to cost restrictions, but he points out that many subsequent trials, including ones that were randomized, have also demonstrated large reductions in complications and mortality following the introduction of the checklist. The list works, he says — as long as it is implemented well. "It turns out to be much more complex that just having the checklist in hand."

TICKING BOXES

Implementation scientists are trying to make sense of that complexity. After the NHS mandated the WHO checklist, researchers at Imperial College London launched a project to monitor the tool's use, and found that staff were often not using it as they should. In a review of nearly 7,000 surgical procedures performed at 5 NHS hospitals, they found that the checklist was used in 97% of cases, but was completed only 62% of the time[8]. When the researchers watched a smaller number of procedures in person, they found that practitioners often failed to give the checks their full attention, and read only two-thirds of the items out loud[9]. In slightly more than 40% of cases, at least one team member was absent during the checks; 10% of the time, the lead surgeon was missing.

Going through all the steps in the list really mattered, the research showed. The more of the checklist that teams completed, the lower the complication rates. Several other studies have also revealed that higher compliance with the checklist is associated with better outcomes.

"If it's used well, if it's used in the original spirit and intention with which it was designed, I think it has real potential," says Sevdalis, who was part of the Imperial College research team. "If it's used for people to tick the box and say, 'Oh yes, we've done it,' but without really thinking about the patient, without really informing their team members about aspects of the procedure that are relevant to them, I don't think the checklist will make any difference."

To find out why checklists were not being used properly, Sevdalis and his colleagues interviewed more than 100 members of operating-theatre staff at 10 NHS hospitals[10]. Half of the respondents reported that senior surgeons and anaesthesiologists sometimes actively resisted the checklists, making it difficult for the rest of the team to complete the tasks. Staff also complained about the checklist itself: that it was poorly worded, time-consuming, inappropriate for certain procedures or redundant with other safety checks. Some also questioned whether there were enough data to support the checklist's use (see 'Why checklists fail').

> "WHEN IT WAS INTRODUCED WITHOUT ANY PROGRAMME OR SUPPORT, IT WAS JUST IMPOSSIBLE, I THINK, FOR TEAMS TO BUY INTO IT."

About one-quarter of the respondents objected to how the checklist had been introduced. Although some hospitals provided training and solicited feedback from staff, at other institutions there was little involvement from those actually working in the operating theatre. That strategy might make it difficult for staff to feel invested in the checklist, and ultimately undermine its correct use. "When it was introduced without any programme or support, it was just impossible, I think, for teams to buy into it," says psychologist Stephanie Russ, who was part of the research team and is now at the University of Aberdeen, UK.

Mary Dixon-Woods, a medical sociologist at the University of Leicester, UK, interviewed staff members at 17 of the ICUs participating in Matching Michigan[11]. She found that by the time the programme began, British hospitals had already been involved in numerous government-led efforts to reduce infections. The checklist, she says, was viewed as "yet another example of these top-down, intrusive, imposed initiatives". It became "something that had to be endured rather than enjoyed". In Michigan, by contrast, the tool was considered new and exciting. And it was not imposed by the government — it was organized by the well-regarded state hospital association, and participation was voluntary.

Dixon-Woods did identify one exemplary ICU, in which a high infection rate fell to zero after Matching Michigan began. The unit was led by a charismatic physician who championed the checklist and rallied others around it. "He formed coalitions with his colleagues so everyone was singing the same tune, and they just committed as a whole unit to getting this problem under control," says Dixon-Woods.

Other work has also found that it might be helpful to enlist local champions who can promote an intervention within a hospital, and some have hinted at how to get colleagues on board. In a 2011 study[12] of five hospitals in Washington state, Gawande and his colleagues found that it is crucial that leaders take the time to explain how to use the checklist and why it should be used. "That might have included pulling on somebody's heart strings, it might have included sharing as much evidence as possible, it might have included talking through the theoretical story or giving some important example," says Sara Singer, a health-policy researcher at the Harvard T. H. Chan School of Public Health in Boston, Massachusetts, who co-authored the study.

WHY CHECKLISTS FAIL

Operating-theatre staff at ten UK hospitals were interviewed about the barriers to implementing the World Health Organization surgical checklist. The biggest problems were:

Staff resisted or failed to complete the checklist.

51%

"When the surgeons weren't on board you were told to 'Oh shut up and let's get on with it.'"

The checklist was inappropriate or illogical.

34%

"It's a bit bizarre and there's a sense of, I'm not actually progressing the patient care with this question."

The checklist was thought to waste time.

29%

"Yet more delay! Oh gosh, we're going to get less work done for the patients."

SOURCE: REF. 10

A LOCAL LIST

Experts also recommend that hospitals modify standard checklists to help the tool fit into the local workflow and to produce a feeling of investment and ownership. Pronovost encouraged the ICUs that participated in the Keystone project to make his checklist their own. "They were 95% the same, but that 5% made it work for them," he says. "Every one of these hospitals thought that theirs was the best."

Pronovost and Dixon-Woods also think that several other factors contributed to the success in Michigan ICUs. Providing the hospitals with regular feedback on their infection rates created social pressure for improvement, they say, and regular in-person workshops allowed staff from different hospitals to share their experiences and created the sense of a shared mission.

Beyond that, logistics are crucial. When Pronovost was first developing his checklist at Johns Hopkins, he noticed that ICU doctors had to go to eight different places to collect all the supplies they needed to perform a sterile central-line insertion. As part of the Keystone programme, hospitals assembled carts that contained all the necessary supplies.

In a 2013 study[13], Dixon-Woods found that an African hospital using the WHO surgical checklist had regular shortages of the basic tools — such as surgical markers, antibiotics and pulse oximeters — that are required to complete the list. But the staff often ticked those boxes anyway; as one anaesthetist pointed out, it was often better for a patient to undergo surgery without these supplies than not to have surgery at all. If the checklist is going to succeed in low-income settings, these problems have to be addressed. "There's no point in having an item that says, 'Have the antibiotics been given?' if there are no antibiotics in the hospital," says Dixon-Woods.

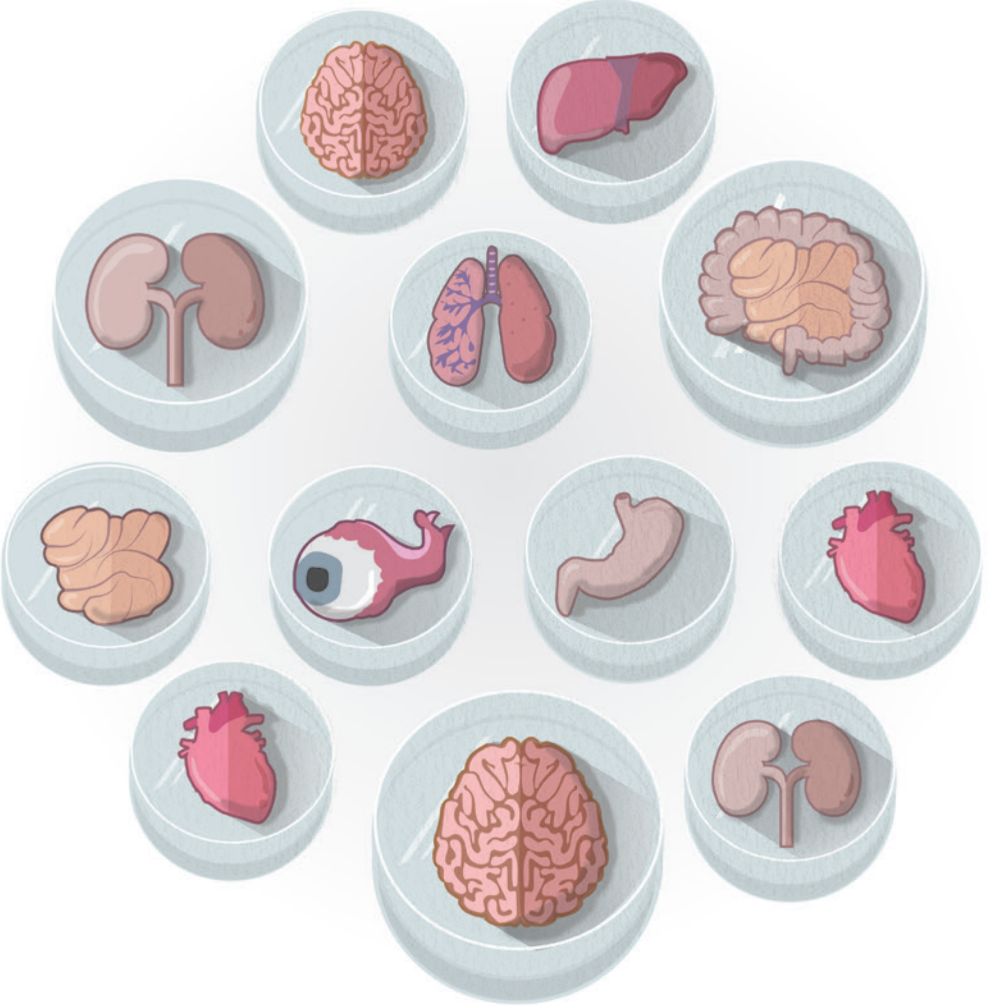
The clear lesson for hospital leaders is that they cannot just dump a stack of checklists in an operating room — they must observe them being used. Are team members all present? Are they rushing, or skipping steps? If so, then the lapses should be discussed and addressed.

Implementation researchers say that the checklist story may hold lessons for the introduction of other programmes in fields including medicine, education and social work. "We have this massive influx of money to develop innovations," says Dean Fixsen, who co-founded the US National Implementation Research Network at the University of North Carolina at Chapel Hill. "But the track record of getting that science into practice where it actually produces the kinds of outcomes that we want to see — that track record is abysmal." Over the past few decades, researchers have published countless papers on evidence-based literacy programmes and teaching strategies. And yet literacy rates for US nine-year-olds, for instance, have barely budged.

Fortunately, Fixsen says, the lessons of implementation science are "completely generalizable", and all programmes could benefit by noting the importance of engaged leadership, local adaptation and user buy-in. "It doesn't matter how good the innovation is, it doesn't matter how much has been invested," says Fixsen. "If we don't have the implementation savvy, we're going to get the crummy outcomes that we have seen decade after decade." ■

Emily Anthes *is a freelance journalist in New York City.*

1. Haynes, A. B. *et al. N. Engl. J. Med.* **360,** 491–499 (2009).
2. Urbach, D. R., Govindarajan, A., Saskin, R., Wilton, A. S. & Baxter, N. N. *N. Engl. J. Med.* **370,** 1029–1038 (2014).
3. Berenholtz, S. M. *et al. Crit. Care Med.* **32,** 2014–2020 (2004).
4. Pronovost, P. *et al. N. Engl. J. Med.* **355,** 2725–2732 (2006).
5. Gawande, A. 'The checklist' *The New Yorker* (10 December 2007); available at go.nature.com/vclrt4
6. Reames, B. N., Krell, R. W., Campbell, D. A. Jr & Dimick, J. B. *JAMA Surg.* **150,** 208–215 (2015).
7. Bion, J. *et al. BMJ Qual. Saf.* **22,** 110–123 (2013).
8. Mayer, E. K. *et al. Ann. Surg.* http://dx.doi.org/10.1097/SLA.0000000000001185 (2015).
9. Russ, S. *et al. J. Am. Coll. Surg.* **220,** 1–11.e4 (2015).
10. Russ, S. J. *et al. Ann. Surg.* **261,** 81–91 (2015).
11. Dixon-Woods, M., Leslie, M., Tarrant, C. & Bion, J. *Implement. Sci.* **8,** 70 (2013).
12. Conley, D. M., Singer, S. J., Edmondson, L., Berry, W. R. & Gawande, A. A. *J. Am. Coll. Surg.* **212,** 873–879 (2011).
13. Aveling, E., McCulloch, P. & Dixon-Woods, M. *BMJ Open* **3,** e003039 (2013).



RISE OF THE ORGANOIDS

Biologists are building banks of mini-organs, and learning a lot about human development on the way.

BY CASSANDRA WILLYARD

ILLUSTRATIONS BY CLAIRE WELSH/*NATURE*

It was an otherwise normal day in November when Madeline Lancaster realized that she had accidentally grown a brain. For weeks, she had been trying to get human embryonic stem cells to form neural rosettes, clusters of cells that can become many different types of neuron. But for some reason her cells refused to stick to the bottom of the culture plate. Instead they floated, forming strange, milky-looking spheres.

"I didn't really know what they were," says Lancaster, who was then a postdoc at the Institute of Molecular Biotechnology in Vienna. That day in 2011, however, she spotted an odd dot of pigment in one of her spheres. Looking under the microscope, she realized that it was the dark cells of a developing retina, an outgrowth of the developing brain. And when she sliced one of the balls open, she could pick out a variety of neurons. Lancaster realized that the cells had assembled themselves into something unmistakably like an embryonic brain, and she went straight to her adviser, stem-cell biologist Jürgen Knoblich, with the news. "I've got something amazing," she told him. "You've got to see it."

Lancaster and her colleagues were not the first to grow a brain in a dish. In 2008, researchers in Japan reported[1] that they had prompted embryonic stem cells from mice and humans to form layered balls reminiscent of a cerebral cortex. Since then, efforts to grow stem cells into rudimentary organs have taken off. Using carefully timed chemical cues, researchers around the world have produced three-dimensional structures that resemble tissue from the eye, gut, liver, kidney, pancreas, prostate, lung, stomach and breast. These bits of tissue, called organoids because they mimic some of the structure and function of real organs, are furthering knowledge of human development, serving as disease models and drug-screening platforms, and might

➲ NATURE.COM
To hear a podcast on organoids, go to:
go.nature.com/xjq1jq

eventually be used to rescue damaged organs (see 'The organoid bank'). "It's probably the most significant development in the stem-cell field in the last five or six years," says Austin Smith, director of the Wellcome Trust/MRC Stem Cell Institute at the University of Cambridge, UK.

The current crop of organoids isn't perfect. Some lack key cell types; others imitate only the earliest stages of organ development or vary from batch to batch. So researchers are toiling to refine their organoids — to make them more complex, more mature and more reproducible. Still, biologists have been amazed at how little encouragement cells need to self-assemble into elaborate structures. "It doesn't require any super-sophisticated bioengineering," says Knoblich. "We just let the cells do what they want to do, and they make a brain."

GROWING A GUT

This shouldn't come as a major surprise, says molecular biologist Melissa Little at the University of Queensland, Australia. "The embryo itself is incredibly able to self-organize; it doesn't need a template or a map." That has been known since the early 1900s, when embryologists showed that sponges that had been broken up into single cells could reassemble themselves. But such work fell out of fashion, and modern biologists have focused their attention on purifying cells and growing them in culture — often in flat layers that do little to mimic normal human tissue.

Studying these cells to understand how an organ functions is like studying a pile of bricks to understand the function of a house, says Mina Bissell, a cancer researcher at the Lawrence Berkeley National Laboratory in California. "We should just begin to make the house," she says. Bissell's work on cultures of breast cells helped to propagate the idea that cells behave differently in 3D cultures than in conventional flat ones. By the mid-2000s, the idea was catching on. The burst of enthusiasm was fuelled by Yoshiki Sasai, a stem-cell biologist at the RIKEN Center for Developmental Biology in Kobe, Japan, who turned heads when he grew a cerebral cortex[1], followed by a rudimentary optic cup[2] and pituitary gland[3] (see *Nature* **488,** 444–446; 2012).

Just a year after Sasai announced his layered cortex, Hans Clevers, a stem-cell researcher at the Hubrecht Institute in Utrecht, the Netherlands, reported the creation of a mini-gut[4]. The breakthrough stemmed from a discovery in 2007, when Clevers and his colleagues had identified intestinal stem cells in mice. In the body, these cells seemed to have an unlimited capacity to divide and replenish the intestinal lining, and one of Clevers' postdocs, Toshiro Sato, was tasked with culturing them in the lab.

Rather than growing the cells flat, the pair decided to embed them in matrigel, a soft jelly that resembles the extracellular matrix, the mesh of molecules that surrounds cells. "We were just trying things," Clevers says. "We hoped that we would make maybe a sphere or a blob of cells." Several months later, when Clevers put his eye to Sato's microscope, he saw more than blobs. The cells had divided, differentiated into multiple types, and formed hollow spheres that were dotted with knobby protrusions. Inside, the team found structures that resembled the intestine's nutrient-absorbing villi as well as the deep valleys between them called crypts. "The structures, to our total astonishment, looked like real guts," Clevers says. "They were beautiful."

The mini-guts, reported in 2009, may prove to be a powerful tool in personalized medicine. Clevers and his team are using them to study the effectiveness of drugs in people with cystic fibrosis, who have genetic defects that affect ion channels and disrupt the movement of water in and out of the cells lining the lungs and intestine. The researchers take rectal biopsies from people with the disease, use the cells to create personalized gut organoids and then apply a potential drug. If the treatment opens the ion channels, then water can flow inwards and the gut organoids swell up. "It's a black-and-white assay," Clevers says, one that could prove quicker and cheaper than trying drugs in people to see whether they work.

THE ORGANOID BANK

Since the late 2000s, biologists have grown a wide variety of rudimentary organs to understand development and for medical uses.

Organoid	Potential application
Cerebral cortex	Understand brain development, as well as neurodegenerative diseases and other disorders
Intestine	Personalized organoids for identifying patient-tailored drugs
Optic cup	Source of retinal tissue for eye therapies
Pituitary gland	Source of therapeutic cells for endocrine disorders
Kidney	Toxicity testing and a source of tissue for transplantation
Liver	Repair of damaged liver
Pancreas	Treat diabetes and identify drugs for pancreatic cancer
Neural tube	Study nerve development and a source of cell therapies
Stomach	Understand stomach development and model gastric disorders such as ulcers
Prostate	Predict effective drug combinations for prostate cancer
Breast	Understand tumour development
Heart	Study cardiac development and how drugs affect it
Lung	Model for lung development, maturation and disease

He has already used the system to assess whether a drug called Kalydeco (ivacaftor), and 5 other cystic-fibrosis drugs, will work in about 100 patients; at least 2 of them are now taking Kalydeco as a result.

Organoids may also help physicians to choose the best therapies for people with cancer. Earlier this year, Clevers revealed that he had grown a bank of organoids from cells extracted from colorectal tumours[5], and David Tuveson, a cancer researcher at Cold Spring Harbor Laboratory in New York, worked with Clevers to generate pancreas organoids using biopsies taken from people with pancreatic cancer[6]. In both cases, the organoids could be used to find drugs that work best on particular tumours. "What patients are looking for is a logical approach to their cancer," Tuveson says. "I'm very excited about what we're learning."

THE SMALL-SCALE STOMACH

That excitement is shared by developmental biologist James Wells, who last year reported that he and his team had created an organoid that resembled part of a human stomach[7].

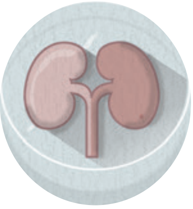
Wells started with a different raw material to Clevers, whose organoids arise from adult stem cells that can generate only a limited number of cell types. Wells, who is at the Cincinnati Children's Hospital Medical Center in Ohio, and his colleagues craft organoids from embryonic stem cells, which have the ability to become almost any type of cell. As a result, they have been able to create mini-organs that are more complex.

A decade ago, Wells and his colleagues began trying to coax human embryonic stem cells to form intestinal cells. When the team manipulated two key signalling pathways, the layer of cells produced tiny round buds. Wells noticed that these 'spheroids' mimicked sections of the primitive gut tube, which forms four weeks after conception. This was thrilling, because he realized that he now had a starting point from which to develop a variety of organoids. "Every organ from your mouth down to your anus — oesophagus, lungs, trachea, stomach, pancreas, liver, intestine, bladder — all of them come from this very primitive tube," he says.

Wells and his colleagues mined the literature and their own experience to determine what chemical cues might send these gut tubes down the developmental path toward a specific organ. Using this strategy, in 2011 the team developed its first human organoid[8], an intestine about the size of a sesame seed. But growing a stomach was a bigger challenge. In humans, the organ has two key areas: the fundus at the top, which churns out acid, and the antrum towards the base, which produces many key digestive hormones — and the signalling pathways that lead to one versus the other were unknown. What is more, "the human stomach is different from the stomachs of most animals that we use in the lab", so there is no good animal model, says Kyle McCracken, a former graduate student of Wells and now a medical student at the centre.

The researchers went for a trial-and-error approach: they made some educated guesses and painstakingly tested different combinations of growth factors. Eventually, the effort paid off. In a 2014 paper[7], Wells and his team revealed that they had created organoids that resembled the antrum. Using these as a model system, the team says that it has figured out the chemical trigger that prompts the development of a fundus. Now the researchers are working to answer other basic questions about stomach development and physiology, such as which factors regulate acid secretion, and they are trying to generate other mini-organs from their primitive gut tubes.

This newfound ability to examine human development excites Daniel St Johnston, a developmental geneticist at the University of Cambridge's Gurdon Institute. "You can actually watch how the cells organize themselves to make complicated structures," he says — something that is impossible in a human embryo. But most organoids are still single tissues, which limits what developmental biologists can learn, he says. "There are certain questions you can't really address because they depend upon the physiology of the whole organism."



THE BABY KIDNEY

Melissa Little has spent more than a decade marvelling at the complexity of the kidney. "It has, in an adult, probably 25–30 different cell types, each doing different jobs," she says. Tubular structures called nephrons filter fluid from the blood and produce urine. The surrounding space, called the interstitium, holds an intricate network of blood vessels and the plumbing that carries urine away.

In 2010, Little and her colleagues started trying to turn embryonic stem cells into a progenitor cell that gives rise to nephrons. For three years, they tried various combinations and timings of growth factors. "It really took a lot of mucking around to make progress," she says. But finally, in 2013, the team landed on just the right mixture. Little had been aiming to produce just the progenitor cells. But when she looked in the dish she saw two cell types spontaneously patterning themselves as they would in an embryo. "There was a moment of, 'Oh wow. Isn't that amazing,'" she says.

This organoid resembles an embryonic kidney rather than an adult one: it has a mix of nephron progenitors and the cells that give rise to urine-collecting ducts[9]. "If you want to get them to mature further, that's where the challenge really lies," Little says. So her team has been working to grow a more-sophisticated version — with blood vessels and interstitium. The hope then is to transplant the mini-organs into mice to see if they will mature and produce urine. "I'm pretty excited about what we can build," Little says.

Because the kidney plays a key part in drug metabolism and excretion, Little thinks that her mini-kidneys could be useful for testing drug candidates for toxicity before they reach clinical trials. And researchers say that other human organoids, such as heart and liver, could similarly be used to screen drug candidates for toxic effects — offering a better read-out on the response of an organ than is possible with standard tissue culture or animal testing.

But Michael Shen, a stem-cell researcher at Columbia University in New York who has created a prostate organoid, is sceptical that these model systems could completely replace lab animals. Animals can show how a therapy affects the immune system, for example, something that organoid systems cannot currently do. "You want to be able to validate your experimental findings in an *in vivo* system," he says. "I view that as a rigorous test."

LITTLE LIVERS

Takanori Takebe was inspired to grow a liver after a chilling spell in New York. While working in the organ-transplantation division at Columbia University in 2010, Takebe saw people die from liver failure owing to a lack of organs. "That was a sad situation," he says. When he looked into tissue engineering, he thought that the usual methods — seeding cells onto an artificial scaffold — seemed destined to fail. Part of the problem, he says, is that adult liver cells are very difficult to grow. "We cannot maintain it in culture for even a couple of hours."

Takebe, who took up a research position at Yokohama City University in Japan, decided to work on induced pluripotent stem (iPS) cells, adult cells that have been reprogrammed to behave like embryonic stem cells. He coaxed human iPS cells into forming liver-cell precursors, or hepatoblasts. In the embryo, hepatoblasts rely on a complex symphony of signals from other nearby cells to mature, and Takebe suspected that these support cells would also be necessary to develop a liver in a dish. He and his colleagues mixed hepatoblasts with such cells — called mesenchymal and endothelial cells — and it worked. The team managed to create 'liver buds', structures no bigger than a lentil that resemble the liver of a six-week-old human embryo[10]. The researchers went on to find that, unlike mature liver cells, such structures can survive in culture for as long as two months.

A liver bud is still a far cry from an entire liver — a hefty, multi-lobed organ composed of tens of billions of hepatocytes. But Takebe hopes that if he can infuse many thousands of buds into a failing organ, he might be able to rescue enough of its function to make a transplant unnecessary. The process seems to work in mice. When Takebe and his group transplanted a dozen of the buds into mouse abdomens, they saw dramatic effects. Within two days, the buds had connected up with the mouse's blood supply, and the cells went on to develop into mature liver cells that were able to make liver-specific proteins and to metabolize drugs. To mimic liver failure, the team wiped out the animals' natural liver function with a toxic drug. After a month, most of the control mice had died, but most of those that received liver bud transplants had survived.

Takebe and his team hope to start human trials in four years. "We will target the children that critically need a liver transplant," he says. He and his colleagues are currently working to make the liver buds smaller and produce them in huge quantities that they can infuse through the large portal vein that feeds the liver. Takebe thinks that the timeline is "doable". But Smith says that the process seems rushed, and that the basic biology of these organs needs to be well understood before they are used in the clinic. "It's like running before you can walk," he says.

Biologists know that their mini-organs are still a crude mimic of their life-sized counterparts. But that gives them something to aim for, says Anthony Atala, director of the Wake Forest Institute for Regenerative Medicine in Winston-Salem, North Carolina. "The long-term goal is that you will be able to replicate more and more of the functionality of a human organ." Already, the field has brought together developmental biologists, stem-cell biologists and clinical scientists. Now the aim is to build more-elaborate organs — ones that are larger and that integrate more cell types.

And Wells says that even today's rudimentary organoids are facilitating discoveries that would have been difficult to make in an animal model, in which the molecular signals are hard to manipulate. "In a Petri dish it's easy," he says. "We have chemicals and proteins that we can just dump onto these cells." ■

Cassandra Willyard *is a science writer based in Madison, Wisconsin.*

1. Eiraku, M. *et al. Cell Stem Cell* **3,** 519–532 (2008).
2. Eiraku, M. *et al. Nature* **472,** 51–56 (2011).
3. Suga, H. *et al. Nature* **480,** 57–62 (2011).
4. Sato, T. *et al. Nature* **459,** 262–265 (2009).
5. van de Wetering, M. *et al. Cell* **161,** 933–945 (2015).
6. Boj, S. F. *et al. Cell* **160,** 324–338 (2015).
7. McCracken, K. W. *et al. Nature* **516,** 400–404 (2014).
8. Spence, J. R. *et al. Nature* **470,** 105–109 (2011).
9. Takasato, M. *et al. Nature Cell Biol.* **16,** 118–126 (2014).
10. Takebe, T. *et al. Nature* **499,** 481–484 (2013).

COMMENT

FROM 1944 *TECHNIQUE*/MIT MUSEUM

Members of the US Army attend a physics lecture at the Massachusetts Institute of Technology in 1944.

From blackboards to bombs

Seventy years after the destruction of Hiroshima and Nagasaki by nuclear weapons, **David Kaiser** investigates the legacy of 'the physicists' war'.

Seventy years ago, on 6 and 9 August 1945, mushroom clouds erupted skyward above the smouldering cities of Hiroshima and Nagasaki, Japan. For the first time — and, so far, the only time — nuclear weapons had been used in combat. Hundreds of thousands of people perished. Many died from the immediate force and fire of the blasts; others succumbed later to acute radiation sickness. Days after the bombs were dropped, Japan surrendered and the Second World War lumbered to a close.

The Second World War marked an unprecedented mobilization of scientists and engineers, and a turning point in the relationship between research and the state. By the end of the war, the nuclear weapons project, code-named the Manhattan Engineer District, absorbed thousands of researchers and billions of dollars. It sprawled across 30 facilities throughout the United States and Canada, with British teams working alongside Americans and Canadians. Allied efforts on radar swelled to comparable scale.

The drama with which the war ended — the detonation of nuclear weapons over cities — cemented the association of the Second World War as 'the physicists' war.' Yet the term had been coined long before August 1945, and originally it had nothing to do with bombs or radar. Rather, the physicists' war had referred to an urgent, ambitious training mission: to teach elementary physics to as many enlisted men as possible.

Both views of how scientists could serve their nations — the quotidian and the cataclysmic — have shaped scientific research and higher education to this day.

CATCHY PHRASE

In late November 1941, just weeks before the United States entered the global conflict, James Conant explained in a newsletter of the American Chemical Society that "this is a physicist's war rather than a chemist's"[1]. Conant was well-placed to know: he was president of Harvard University in Cambridge, Massachusetts, chair of the US National Defense Research Committee (NDRC), and a veteran of earlier chemical-weapons projects.

The phrase had instant appeal; others quickly began to quote it. In 1949, for example, *Life* magazine profiled[2] physicist J. Robert Oppenheimer, who had served as scientific director of the wartime Los Alamos laboratory in New Mexico, a central node of the Manhattan Project. Referring to massive military projects such as the bomb and radar, the reporter invoked "the popular notion" that the Second World War had been "a physicists' war".

By that time, the meaning of Conant's formulation seemed self-evident. The First World War, with its notorious battlefield uses of poison gases such as phosgene and chlorine, had been dubbed the chemists' war. The bomb and radar presented a logical counterpoint.

But Conant had very different ideas when he introduced the now-famous phrase. It was hardly clear in November 1941 that either the bomb or radar would change the tide of the war. The Radiation Laboratory, or 'Rad Lab', at the Massachusetts Institute of Technology (MIT) in Cambridge — which served as headquarters for the Allied effort to improve radar — was just one year old. A prototype radar system had recently been rejected by a US Army review board, and NDRC funding had nearly been revoked. The Manhattan Project did not exist yet. Los Alamos still housed a private boys' school. A year and a half would elapse before the mud-caked ranch houses were requisitioned for the top-secret laboratory.

> *"Physics teachers became a rationed commodity."*

There is also the matter of secrecy. Conant oversaw both radar development and the nascent nuclear-weapons programme; information about each was strictly classified. An experienced, high-ranking government adviser such as Conant surely did not intend to disclose some of the nation's most closely guarded secrets.

And there is the nature of the radar and bomb projects themselves. Although each was directed by physicists, they teemed with specialists of all stripes. By the end of the war, physicists were a minority — only about one-fifth — of the Rad Lab staff. At Los Alamos, the wartime organization chart displayed the groups — metallurgy, chemistry, ballistics, ordnance and electrical engineering, as well as physics — arranged in a circle, connected by spindly links. No group appeared on top directing the others. Researchers at both the Rad Lab and Los Alamos forged new kinds of hybrid, interdisciplinary spaces during the war. Neither could be categorized simply as a physics laboratory[3] (see D. Kaiser *Nature* **505,** 153–155; 2014).

So what was Conant talking about?

CLASSROOM MOBILIZATION

To most scientists and policy-makers in the early 1940s, the physicists' war referred to a massive educational mission.

In January 1942, the director of the American Institute of Physics (AIP), Henry Barton, citing Conant, began issuing bulletins entitled 'A Physicist's War'. Barton reasoned that "the conditions under which physicists can render services to their country are changing so rapidly" that department chairs and heads of laboratories needed some means of keeping abreast of evolving policies and priorities. The monthly bulletins focused on two main topics: how to secure draft deferments for physics students and personnel, and how academic departments could meet the sudden demand for more physics instruction.

Modern warfare, it seemed, required rudimentary knowledge of optics and acoustics, radio and circuits. Before the war, the US Army and Navy had trained technical specialists from within their own ranks, at their own facilities. The sudden entry of the United States into the war required new tactics. University physicists, consulting with army and navy officials, reported early in the conflict that enrolments in high-school physics classes would need to jump by 250%. Their goal: half of all high-school boys in the country should spend at least one class per day focusing on electricity, circuits and radio[4].

The navy and the army also called for massive numbers of military personnel to receive basic training in physics at established colleges and universities. Draft curricula circulated between military officials and the AIP. The army, for example, wanted the new courses to emphasize how to measure lengths, angles, air temperature, barometric pressure, relative humidity, electric current and voltage. Lessons in geometrical optics would emphasize applications to battlefield scopes; lessons in acoustics would drop examples from music in favour of depth sounding and sound ranging.

So acute was the need to teach elementary physics that a special committee recommended that university departments discontinue courses in atomic and nuclear physics for the duration of the war so as to devote more teaching resources to truly "essential" material[5].

Between December 1942 and August 1945, a quarter of a million students passed through elementary physics classes at US colleges and universities. Staffing the inflated classrooms required military-style planning and logistics. Barton's bulletins warned that any universities found to be hoarding valuable physics teachers — much less poaching them from other schools — would be subject to "severe criticism". Barton developed a complicated formula for what he termed the acceptable "ratio of genuine to 'ersatz' teachers of physics" in any given institution. Physics teachers became a rationed commodity: like rubber, gasoline and sugar, they were in critically short supply.

Draft policies quickly followed. The US government created a National Committee on Physicists in December 1942 — the first of its kind for any academic speciality — to advise local draft boards on the need for teaching-related deferments. Soon the phrase the physicists' war echoed throughout newspapers, popular magazines and even congressional testimony. Use of the phrase peaked in 1943, long before there was much to report (classified or otherwise) about the Manhattan Project (see 'Branding exercise').

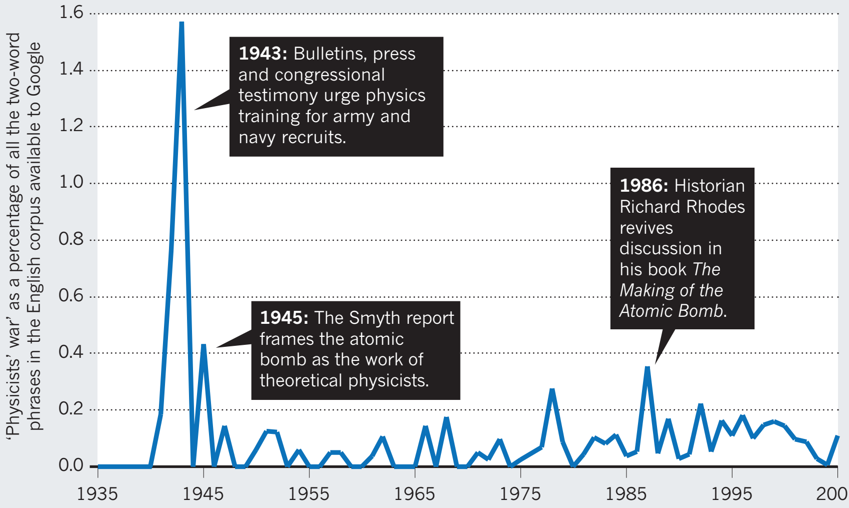
SECRECY AND THE SMYTH REPORT

Use of the phrase the physicists' war rebounded every decade or so, usually around an anniversary of the bombings of Hiroshima and Nagasaki. The highest post-war peak accompanied the publication of Richard Rhodes's Pulitzer-prizewinning book, *The Making of the Atomic Bomb*, in 1986. By then, Conant's phrase had long since been linked with classified military projects rather than classroom instruction.

The transition began almost as soon as the bombs were dropped on Japan. General Leslie Groves, who was overseeing the

BRANDING EXERCISE

The phrase 'the physicists' war', coined to call for mass education as the United States entered the Second World War, took on a new life after August 1945.



1943: Bulletins, press and congressional testimony urge physics training for army and navy recruits.

1945: The Smyth report frames the atomic bomb as the work of theoretical physicists.

1986: Historian Richard Rhodes revives discussion in his book *The Making of the Atomic Bomb.*

SOURCE: GOOGLE

MIT MUSEUM

Reserve Officer Training Corps students use a portable radio as part of their training at MIT (about 1944).

Manhattan Project, anticipated that the government would need to have some information ready to release about the top-secret nuclear-weapons project — pre-cleared and available for wide distribution — in case the bombs were ever used. Early in the project, he tapped nuclear physicist Henry DeWolf Smyth at Princeton University in New Jersey to spend the war visiting each Manhattan Project site, compiling a technical report that would be suitable for public dissemination.

On the evening of 11 August 1945, just two days after the bombing of Nagasaki, the US government released Smyth's 200-page document under the ponderous title, 'A General Account of Methods of Using Atomic Energy for Military Purposes under the Auspices of the United States Government, 1940–1945'. Quickly dubbed 'the Smyth report', copies flew off the shelves. The original Government Printing Office edition ran out so quickly that Princeton University Press published its own edition late in 1945, under the more manageable title, *Atomic Energy for Military Purposes*, which sold more than 100,000 copies in a year.

Security considerations dominated what Smyth could include. Only information that was already widely known to working scientists and engineers, or which had "no real bearing on the production of atomic bombs", was deemed fit for release. Little of the messy combination of chemistry, metallurgy, engineering or industrial-scale manufacturing met these criteria; these aspects of the huge project, crucial to the actual design and production of nuclear weapons, remained closely guarded[6].

So Smyth focused on ideas from physics, pushing theoretical physics, in particular, to the forefront. Ironically, most people read in Smyth's report the lesson that physicists had built the bomb (and, by implication, had won the war)[6,7]. Later reports, such as 'Essential Information on Atomic Energy', issued in 1946 by the new Special Committee on Atomic Energy of the US Senate, borrowed liberally from the Smyth report, depicting nuclear weapons as the latest in a series of developments in theoretical physics. A chronological table at the end extended the narrative as far back as 400 BC to the ancient Greek atomists. There was little mention of the Berlin chemistry laboratory in which nuclear fission had been discovered late in 1939, much less the work of chemical engineers at US company DuPont, who scaled up plutonium-producing nuclear reactors during the war.

LONG SHADOW

The change in referent for the physicists' war — from blackboards to bombs — had serious implications. After the war, physicists in the United States bore the largest brunt of any academic group during the 'red scare' of the 1950s, promoted by Senator Joseph McCarthy. The House Un-American Activities Committee held 27 hearings on allegations against physicists, twice the number for any other scholarly discipline. If nuclear weapons had been made by physicists, the reasoning went, then physicists must have special access to the 'atomic secrets' with which such bombs could be made. Thus the loyalties of this group required the closest scrutiny[8].

Meanwhile, the two meanings of the physicists' war blurred together as the cold war intensified. More and more universities became contracting sites for military and defence agencies, continuing the model that Conant and others had forged during the war. Physicists' research budgets ballooned, and enrolments grew faster than in any other field, doubling every few years.

More physicists were trained in the United States, United Kingdom and Soviet Union in the quarter-century after the war than had been trained throughout all of previous history. Yet the aims of the training shifted in the 1950s and 1960s. Rather than teaching soldiers some elementary physics to prepare them for the battlefield, US officials spoke of creating a 'standing army' of physicists, who could work on nuclear-weapons projects without delay should the cold war ever turn hot[9].

Three decades after 1945, years into the slog of the Vietnam War, many critics grew uneasy with the close association between physics and war. Campus protesters demanded that the defence department get out of the higher-education business. At universities across the United States, physicists' laboratories became frequent targets for sit-ins and even Molotov cocktails[10].

After the protesters dispersed and the tear gas lifted, several things had become clear. 'The physicists' war' had massively altered the structure of the US university system, the organization of scientific research, and the relationship between national defence and higher education. ■

David Kaiser *is professor of physics and of the history of science at the Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. He is author of* How the Hippies Saved Physics: Science, Counterculture, and the Quantum Revival.
e-mail: dikaiser@mit.edu

1. Conant, J. B. *Chem. Eng. News* **19,** 1237–1238 (1941).
2. Barnett, L. 'J. Robert Oppenheimer' *Life* 120–138 (10 October 1949).
3. Galison, P. *Image and Logic: A Material Culture of Microphysics* (Univ. Chicago Press, 1997).
4. Havighurst, R. J. & Lark-Horovitz, K. *Am. J. Phys.* **11,** 103–108 (1943).
5. Cope, T. D. *et al. Am. J. Phys.* **10,** 266–268 (1942).
6. Schwartz, R. P. *The Making of the History of the Atomic Bomb: The Smyth Report and the Historiography of the Manhattan Project* PhD thesis, Princeton Univ. (2008).
7. Gordin, M. *Five Days in August: How World War II Became a Nuclear War* (Princeton Univ. Press, 2007).
8. Kaiser, D. *Representations* **90,** 28–60 (2005).
9. Kaiser, D. *Hist. Stud. Phys. Sci.* **33,** 131–159 (2002).
10. Moore, K. *Disrupting Science: Social Movements, American Scientists, and the Politics of the Military, 1945–1975* (Princeton Univ. Press, 2008).

UDO WEITZ/BLOOMBERG/GETTY

Mountain pine beetles have destroyed hundreds of thousands of hectares of forest in Canada.

Define biomass sustainability

The future of the bioeconomy requires global agreement on metrics and the creation of a dispute resolution centre, say **Roeland Bosch**, **Mattheüs van de Pol** and **Jim Philp**.

The bioeconomy is rising up the political agenda. More than 30 countries have announced that they will boost production of renewable resources from biological materials and convert them into products such as food, animal feed and bioenergy. Non-food crops, such as switchgrass (*Panicum virgatum*), are the main focus, as well as agricultural and forestry residues and waste materials and gases.

It is one thing to write a report; it is another to put a plan into action sustainably. The biggest conundrum is reconciling the conflicting needs of agriculture and industry. In a post-fossil-fuel world, an increasing proportion of chemicals, plastics, textiles, fuels and electricity will have to come from biomass, which takes up land. By 2050, the world will also need to produce 50–70% more food[1], increasingly under drought conditions and on poor soils.

There is no consensus on what 'sustainable' means. Biomass assessment is a patchwork of voluntary standards and regulations. With many schemes comes a lack of comparability. Confusion leads to mistrust and protectionism, international disputes and barriers, slow investment and slower growth.

For example, greater use of wood for electricity generation or heating may decrease greenhouse-gas emissions if it displaces coal. But retaining forests also sequesters carbon and protects biodiversity. Increased demand boosts wood-pellet prices, and puts pressure on businesses, such as saw mills, that use wood. The balance of who saves or creates emissions shifts when biomass is exported.

The geopolitical implications mirror those of crude oil. Developed countries that lack fossil fuels are thirsty for renewable energies. Some developing countries may be tempted to meet that demand without accounting for the environmental or social cost. It is in everyone's interests to harmonize sustainability standards and head off disputes before they arise. Governments should agree on criteria and define metrics for assessing biomass sustainability. And they should consider creating a centre for resolving disputes that arise over competition for land and biomass.

NO CONSENSUS

In 2012, the United States and the European Union laid down their intentions to grow their bioeconomies[2,3]. Now, the G7 industrialized countries[4] and at least 20 others either have a dedicated bioeconomy strategy in place (including Finland, Malaysia and South Africa) or have policies consistent with growing a bioeconomy (including Australia, Brazil, China, India and Russia).

The bioeconomy of Malaysia, for example, is expected to grow by 15% per year to 2030. The palm-oil industry is central to that plan, as it is elsewhere in southeast Asia.

Making up 45% of the world's edible oil, palm oil can also be processed into biodiesel. The oil-palm crop is also more effective at sequestering carbon than other major crops. Using genomics in selective breeding offers great potential for improvements to the economics of palm-oil production[5].

Land disputes between palm-oil companies and local communities have already begun. Between 2006 and 2010, Indonesia's palm-oil plantation area increased dramatically, from 4.1 million to 7.2 million hectares. The increase has been accompanied by a rise in deforestation, water pollution, soil erosion and air pollution, as well as restrictions on traditional land-use rights and land losses, increasing land scarcity and land prices[6].

A situation that arose between Canada and the EU in 2012 illustrates how rational decision-making in different countries can lead to disputes. The EU's Renewable Energy Directive sustainability criteria for biofuels and bioliquids are non-binding for solid biomass. EU biomass sustainability standards also prohibit the use of 'primary forest' materials for bioenergy.

In Canada, forests are deemed sustainable by measures of woodland structure, composition and degree of 'naturalness'. Overall, the area affected by natural disturbances such as insect infestations and wildfire is larger than the total area of logging — and the use of such damaged trees for bioenergy holds potential. But because it stems from primary forest that has not been harvested or regrown, such wood would be excluded from importation into the EU.

A dispute arose in 2012 between an environmental organization and an energy company wishing to ship wood pellets to Europe that had originated from Canadian primary forest infested with the mountain pine beetle (*Dendroctonus ponderosae*). The Dutch government used the case to see whether mediation might work in such circumstances — and it did. The dispute was settled.

Increasing demand for biomass makes it likely that such disputes will recur. Limited land mean that Europe cannot grow enough biomass to meet its own future demand. Depending on bioenergy policies, biomass use is expected to continue to rise to 2030 and imports to Europe are estimated to triple by 2020. Wood-pellet use for large-scale power generation is increasing dramatically in Europe. Some countries including Germany and Denmark have become net importers. Europe may import[7] 80 million tonnes of solid biomass per year by 2020.

Today's biomass situation bears similarities to that in the 1980s, when a system of national agricultural subsidies in Europe threatened to start trade wars. Policies directed at producing more food combined with rapid technical progress and structural changes led to agricultural trade barriers. Domestic surpluses of farm goods were stocked or exported with subsidies — giving rise to the European 'butter mountains' and 'wine lakes' — by protecting farm producers at the cost of domestic consumers and producers abroad. The costs weighed heavily on government budgets. Consumers in countries with protected markets faced higher food bills, and producers in other countries were penalized by restrictions on access to those markets[8].

The Organisation for Economic Co-operation and Development (OECD) helped to resolve that situation by developing standards for agricultural subsidies, which are accepted globally. An analogous, internationally agreed biomass sustainability governance framework is now needed.

SUSTAINABILITY METRIC

A metric for evaluating biomass sustainability needs to be designed. Social as well as environmental and economic factors must be included. As yet there is no consensus on what criteria should be used. For example, international stakeholders (non-governmental organizations, policy-makers, research and development, bioenergy producers, end-users and traders) from 25 European and 9 non-European countries surveyed in 2011 agreed[9] unanimously on only one criterion — minimization of greenhouse-gas emissions.

Aggregation of sustainability issues into a single measure requires complicated trade-offs between, say, kilograms of carbon dioxide emissions and labour conditions. Practitioners' own weightings are subjective. Life-cycle analysis — a technique to assess environmental impacts of all the stages of the manufacture, use and disposal of a product — does not look at social impacts. To define an index, multiple variables must be expressed using a common denominator.

Using price information is understood by policy-makers and the market. But placing monetary values on social and ethical costs and benefits is contentious. Differences between developed and developing countries require careful handling — for example, the different reactions to placing a cost on child labour.

A starting point could be the total factor productivity (TFP) metric used to measure agricultural sustainability. The TFP reflects the rate of transformation of inputs (capital, labour, materials, energy and services) into outputs (biomass stock). A cost is attributed to each and to the negative social and economic impacts.

To develop the TFP approach into an integrated methodology for assessing biomass sustainability, it should take account of changing conditions and local situations in biomass importing and exporting countries. For example, just like oil, the price of biomass will fluctuate. Expensive producers would be hurt. Sustainability assessment will need to take account of such fluctuations. A tenet of a bioeconomy is decentralized feedstock access — biomass-sustainability assessment needs the flexibility to take such matters into account.

Biomass metrics could be aligned with the United Nations' Sustainable Development Goals, the indicators[10] of which are similar. A step towards this alignment may be achieved in November at the Global Bioeconomy Summit in Berlin. This meeting of more than 500 leaders from policy, research, industry and civil society organized by the German Bioeconomy Council (an independent advisory body to the German government) could result in recommendations around global governance and international cooperation.

Beyond that, international agreement is needed on the key biomass sustainability criteria. The OECD could host initial discussions including exploring the setting up of an international biomass dispute settlement facility. Developing countries and biomass producers outside the OECD should also be represented at such talks. These possibilities will be discussed at the OECD Committee for Scientific and Technological Policy ministerial-level meeting in October hosted by the South Korean government in Daejeon. ■

"Differences between developed and developing countries require careful handling."

Roeland Bosch *is the former chief economist at the Unit of Green Growth and Biobased Economy, Ministry of Economic Affairs, The Hague, The Netherlands.* **Mattheüs van de Pol** *is a policy-maker at the Unit of Green Growth and Biobased Economy, Ministry of Economic Affairs, The Hague, The Netherlands.* **Jim Philp** *is a policy analyst at the Organisation for Economic Co-operation and Development, Paris France.*
e-mail: james.philp@oecd.org

1. Food and Agriculture Organization of the United Nations. *The State of Food and Agriculture: Livestock in the Balance* (FAO, 2009).
2. The White House. *National Bioeconomy Blueprint* (2012).
3. European Commission. *Innovating for Sustainable Growth: A Bioeconomy for Europe* (European Commission, 2012).
4. Bioökonomierat. *Bioeconomy Policy: Synopsis and Analysis of Strategies in the G7* (German Bioeconomy Council, 2015).
5. Singh, R. *et al. Nature* **500,** 340–344 (2013).
6. Obidzinski, K., Andriani, R., Komarudin, H. & Andrianto, A. *Ecol. Soc.* **17,** 25 (2012).
7. Scarlat, N., Dallemand, J. F., Motola, V. & Monforti-Ferrario, F. *Renew. Energy* **57,** 448–461 (2013).
8. Organisation for Economic Co-operation and Development. *OECD's Producer Support Estimate and Related Indicators of Agricultural Support* (OECD, 2010).
9. van Dam, J. & Junginger, M. *Energy Policy* **39,** 4051–4066 (2011).
10. Lu, Y., Nakicenovic, N., Visbeck, M. & Stevance A.-S. *Nature* **520,** 432–433 (2015).

The views expressed are those of the authors and not necessarily those of the OECD or of the governments of its member countries.

Correspondence

Many ways to access hominin fossil finds

Michael Cherry's perspective of elitist tourism at South Africa's Cradle of Humankind World Heritage Site in Gauteng does not take into account extensive efforts by scientists, the government and the private sector to bring a wider understanding of human evolution and our shared heritage to the South African public (*Nature* **523,** 33; 2015).

Excavations of hominin fossils at the Malapa caves, viewed from an overhead structure (pictured), may need wealthy tourists to support access to the site, but plans include free school tours on at least one day each month. The Rising Star cave will be open as a nature and heritage reserve for an entry fee of just a few rands (12 rands is US$1). A virtual lab at the Maropeng visitor centre will also allow people to watch Malapa fossils being prepared online.

DARREN ALEXANDER

Last year, 38,000 schoolchildren visited our education facilities at the active palaeontological site of Sterkfontein. The outreach programmes at the University of the Witwatersrand's Evolutionary Studies Institute, together with privately funded activities, reach about 200,000 more schoolchildren, most from disadvantaged backgrounds.

Public exhibitions of our fossil hominin treasures are held frequently at museums around South Africa, with up to 15 on display at any one time. The government and the University of the Witwatersrand have also gifted casts of *Australopithecus sediba* fossils to museums across Africa and around the world.

Lee R. Berger *University of the Witwatersrand, South Africa.*
lee.berger@wits.ac.za

Initiatives to bridge faith and science

Two initiatives aim to increase awareness and acceptance of science by US Christian communities, some of which resist science-education efforts.

BioLogos (http://biologos.org) was founded in 2007 by Francis Collins, then leader of the Human Genome Project, to encourage other Christians to accept evolution in the context of their faith. Trust and respect for Collins has been key to its success. Its grant programme has so far disbursed a total of US$3.9 million to 37 faith–science partnerships.

In fostering dialogue between theologians and scientists who are Christians, BioLogos is forging a middle ground between presentations of science that are antagonistic towards faith and faith that will not accommodate science. Last month, for example, a BioLogos conference of scientists, theologians and pastors helped to articulate the overlap between theology and evolutionary theory (see go.nature.com/ovnpwb).

A programme by the American Association for the Advancement of Science takes a different approach, in partnership with the Association of Theological Schools (www.scienceforseminaries.org). Their pilot project has helped ten seminaries since 2013 to integrate science into the curriculum for training religious leaders. One seminary includes evolutionary biology in courses on biblical interpretation; others teach neuroscience, genetics or ecology within explorations of identity and environmental stewardship.

Science education is a public good that we as scientists should help to reinforce across all faiths with partnerships such as these.

S. Joshua Swamidass *Washington University in St. Louis, Missouri, USA.*
swamidass@wustl.edu

Tree rings track climate trade-offs

The information held in annual tree rings offers further insight into the potential of the world's forests to slow global warming (see *Nature* **523,** 20–22; 2015). These data reveal fluctuations in growth rate caused by climatic, physiological and ecological factors, and provide long-term and spatially extensive records.

Tree-ring measurements provide information on functional trade-offs that can affect a tree's future, such as altered hydraulic and growth responses to the long-term rise in atmospheric carbon dioxide (see P. van der Sleen *et al. Nature Geosci.* **8,** 24–28; 2015). And unlike data from remote sensing of forests and monitoring of tree diameters, tree-ring observations extend over centuries.

Such studies also mitigate the problems you mention of investigating small forest patches, because the size and number of research areas can be adjusted.

A relatively small percentage of ring-forming species in tropical forests and the fading record of tree-ring collections further back in time are impediments. Also, most existing tree-ring collections are not representative of trees that lived in the past (F. Babst *et al. Oecologia* **176,** 307–322; 2014).

We suggest that combining retrospective tree-ring analysis, repeated forest inventory measurements and modelling studies could more effectively predict responses of the world's forests to climate change.

Pieter A. Zuidema *Wageningen University, the Netherlands.*
David Frank *Swiss Federal Research Institute WSL, Birmensdorf, Switzerland.*
david.frank@wsl.ch

Universities aim for a sustainable future

A survey of several European universities shows that corporate responsibility for the economic, social and environmental effects of commercial activities is slowly spreading to the non-profit sector in Europe — including to institutions of higher education.

The survey, undertaken in 2014, polled 73 members of the European University Association, representing 18 countries across a range of institutional age, size and ranking (see Y. Fassin in *Proc. 26th Int. Assoc. Bus. Soc.*; in the press). Of those 73 universities (response rate 11%), 82% had incorporated sustainability and the European Commission's corporate social responsibility recommendations into their development strategies; 70% were already reporting on their initiatives.

The survey revealed that these universities tend to emphasize sustainable development over social responsibility. Perhaps because social responsibility is a given for universities, sustainability for future generations of students seems to hold more sway with university management.

Yves Fassin *Ghent University, Belgium.*
yves.fassin@ugent.be

NEWS & VIEWS

For News & Views online, go to
nature.com/newsandviews

SUSTAINABILITY

Bypassing the methane cycle

A genetically modified rice with more starch in its grains also provides fewer nutrients for methane-producing soil microbes. This dual benefit might help to meet the urgent need for globally sustainable food production. See Letter p.602

PAUL L. E. BODELIER

Despite being much less abundant in the atmosphere than carbon dioxide, methane has a higher capacity to absorb heat emitted from Earth's surface, and thereby contributes substantially to global warming[1]. The demonstration[2] that reducing emissions of this greenhouse gas can be achieved more easily and more rapidly than for CO_2 has spurred an avalanche of studies on possible methane-mitigation strategies. Rice cultivation is the largest single source of methane linked to human activity, and methane emissions from rice paddies are expected to increase with a rising global demand for food. On page 602 of this issue, Su *et al.*[3] describe a 'high-starch, low-methane' rice variety that represents a tremendous opportunity for more-sustainable rice cultivation.

In rice, the leaves and stems take up CO_2, which is transformed through photosynthesis into sugars that are used to produce plant biomass or storage compounds, such as starch, in the shoots, roots and rice grains. Carbon from decaying roots or that is directly released by roots into the soil in the form of sugars, amino acids and organic acids, can be transformed by decomposer microorganisms into substrates (CO_2, hydrogen and acetate). In the absence of oxygen, these substrates are turned into methane by methanogenic microorganisms. The methane can remain in the soil, escape to the atmosphere through diffusion into and emission by the rice plants, or be intercepted by methane-consuming bacteria in the root zone or the soil's surface layer. These bacteria use methane as a substrate for oxygen-requiring respiration, producing CO_2. Hence, methane emission from rice soils is determined by the balance of methane-producing and methane-oxidizing microbes, the availability of other substrates, and the activity of microbes competing for these compounds — collectively, these processes constitute the methane cycle (Fig. 1).

Existing efforts to mitigate rice-associated methane emissions have focused mainly[4] on agricultural practices — such as water management, fertilizer use, tillage and crop selection — that alter the environmental conditions for methanogenic microorganisms. However, these measures are labour intensive and their applicability varies between rice-cropping systems and between countries. In 2002, it was observed[5] that the larger the amount of grain carried by the rice plants, the less methane is emitted, because carbon fixed in the grains does not become available for soil microbes to turn into methane. This observation suggested a globally applicable methane-mitigation solution: produce a rice plant with a higher proportion of its carbon in the stems and grains than current varieties. Crops of this plant would not only result in less methane emissions, but also give higher grain yield and higher nutritional value. But it has taken until now for such a plant to be generated. Su *et al.* present evidence from field and greenhouse trials of varieties of transgenic rice called *SUSIBA2* that fulfil these criteria.

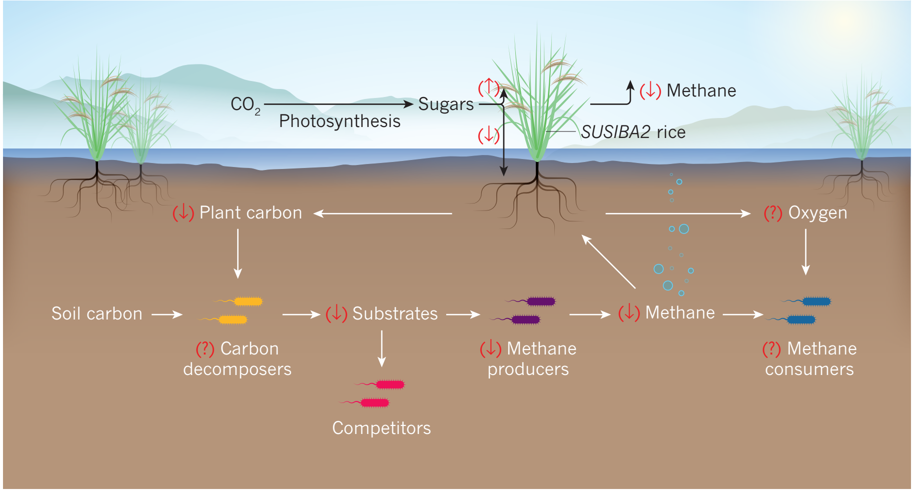


Figure 1 | High-starch rice in the environment. The transgenic *SUSIBA2* rice described by Su *et al.*[3] produces grains with a high starch content by diverting more carbon (derived from photosynthesis) into grains and stems, and less into roots (red arrows). This results in less carbon being input into the soil by the plants and thus being available to decomposer microorganisms that supply carbon-containing substrates to methane-producing microbes. These effects combine to dramatically reduce methane emission from areas planted with this rice strain. However, the amount of methane emitted depends on a complex interplay between plant physiology and the activity of these and other microorganisms, including competitors for substrates and methane consumers. These interactions can all be influenced by the amount and type of carbon compounds and nutrients available, as well as by the amount of oxygen expelled by the roots. The effects of *SUSIBA2* rice on many aspects of this cycle are not yet clear (question marks).

The authors generated their rice varieties by transferring genes from barley that are responsible for the production of starch in stems and grains; in rice, these genes lead to higher starch production, and hence a higher demand for sugars, in these plant parts. The transgenic rice displayed the expected properties of higher seed weight and higher starch content in seeds and stems, but no change in starch levels in leaves and roots and a markedly lower root biomass. By screening the expression of a large set of genes involved in converting sugar into starch, the authors confirmed that the inserted genetic elements were effective only in the seeds and stems.

The 'cherry on top' of these genetic efforts was a significant reduction in methane emitted from *SUSIBA2* rice plants, compared with a widely grown unmodified variety. The methane was measured by covering the plants with transparent gas-collection chambers. A decrease in methane emission was seen in trials of two variants of *SUSIBA2* rice in three regions of China, measured in three consecutive growing seasons. The authors also found fewer methanogenic microorganisms on and around the roots of *SUSIBA2* rice, suggesting that there was less plant-derived carbon-containing substrate available for methanogens.

Although Su and colleagues have made the groundbreaking demonstration that high-starch, low-methane rice plants can be generated, their study raises many issues. The most obvious is that *SUSIBA2* rice is a transgenic plant, and thus raises biological and ethics concerns. In addition to the general questions surrounding the use of genetically modified crops for human consumption, and how access to seed for such crops is controlled, we do not yet have a clear picture of how this modification affects rice plants' survival and general function.

Long-term and frequent measurements of methane emissions from areas planted with normal and transgenic rice are needed to estimate what the annual global effect of the widespread use of this crop would be, and how it compares with that of other methane-mitigation strategies. Even more important will be assessment of the long-term consequences of lower carbon and oxygen input by the roots of *SUSIBA2* plants on soil processes and the microbes that carry them out (Fig. 1). It has recently been shown[6] that highly specific assemblages of microbial species occur in, on and around rice roots, and that not all members use plant-exuded carbon[7]. Long-term reduction of root-exuded carbon might alter the composition of these communities, with unknown consequences for microbes that are plant pathogens or that benefit the plants, such as the bacteria that decompose organic material and deliver essential plant nutrients[8].

To compensate for the possible reduction in plant nutrients, larger amounts of nitrogen fertilizer would need to be applied. This can affect both methane producers and consumers[9] and lead to undesirable environmental effects, such as nitrate leaching to groundwater and emission of the potent greenhouse gas nitrous oxide. Also crucial for the amount of methane emitted is the activity of methane-consuming aerobic bacteria. The oxygen they use flows though the plant stems and roots into the soil by the same route taken by methane moving out of the soil into the atmosphere, and it is not known how the transport of gases is affected in the transgenic rice.

Thus, translocating more carbon to the stems and seeds of *SUSIBA2* rice may bypass methane cycling, but this activity has the potential to affect a multitude of processes involving soil carbon, nutrients and microbial activity, with knock-on effects for the sustainability of rice cultivation. However, Su and colleagues have achieved the feat of making high-starch rice available, and this will spur scientists worldwide to conduct experiments to verify whether this variety will enable more-sustainable cultivation of the crop that feeds half the human population. ■

Paul L. E. Bodelier *is in the Department of Microbial Ecology, Netherlands Institute of Ecology (NIOO-KNAW), 6708 PB Wageningen, the Netherlands.*
e-mail: p.bodelier@nioo.knaw.nl

1. IPCC *Climate Change 2013: The Physical Science Basis* (eds Stocker, T. F. *et al.*) (Cambridge Univ. Press, 2013).
2. Montzka, S. A., Dlugokencky, E. J. & Butler, J. H. *Nature* **476,** 43–50 (2011).
3. Su, J. *et al. Nature* **523,** 602–606 (2015).
4. Hussain, S. *et al. Environ. Sci. Pollut. Res. Int.* **22,** 3342–3360 (2015).
5. Denier van der Gon, H. A. C. *et al. Proc. Natl Acad. Sci. USA* **99,** 12021–12024 (2002).
6. Edwards, J. *et al. Proc. Natl Acad. Sci. USA* **112,** E911–E920 (2015).
7. Hernández, M., Dumont, M. G., Yuan, Q. & Conrad, R. *Appl. Environ. Microbiol.* **81,** 2244–2253 (2015).
8. Philippot, L., Raaijmakers, J. M., Lemanceau, P. & Van der Putten, W. H. *Nature Rev. Microbiol.* **11,** 789–799 (2013).
9. Bodelier, P. L. E. & Steenbergh, A. K. *Curr. Opin. Env. Sustain.* **9–10,** 26–36 (2014).

This article was published online on 22 July 2015.

INORGANIC CHEMISTRY

Movies of a growth mechanism

A microscopy technique has been used to study the formation and growth of crystals of porous solids known as metal–organic frameworks in real time. The findings will aid the design of methods for making these useful compounds.

KRISTA S. WALTON

Materials called metal–organic frameworks (MOFs) have sparked intense interest over the past few decades. In particular, those that form permanently porous architectures have tremendous potential for applications such as chemical sensing, gas storage and catalysis. But techniques for synthesizing these compounds are still often developed through trial and error — in part because the mechanisms that dictate the self-assembly of MOF unit cells from their constituent metal ions and ligands, and their subsequent growth into nanoparticles, are largely unknown and difficult to observe. Writing in the *Journal of the American Chemical Society,* Patterson *et al.*[1] help to solve this problem by reporting the first observations of the crystallization of MOFs made in real time, using a technique called liquid-cell transmission electron microscopy.

Previous studies of MOF crystallization have made use of various *ex situ* and *in situ* analytical techniques, including high-resolution transmission electron microscopy (HRTEM) and energy-dispersive X-ray diffraction (EDXRD). For example, HRTEM has been used to examine the crystallization of the well-characterized MOF-5, by analysing samples taken at various time intervals early in the compound's synthesis[2]. Time-resolved *in situ* EDXRD has been used to determine the kinetics of MOF crystallization as a function of parameters that included pH, temperature and ligand length[3,4]. But the challenge of observing

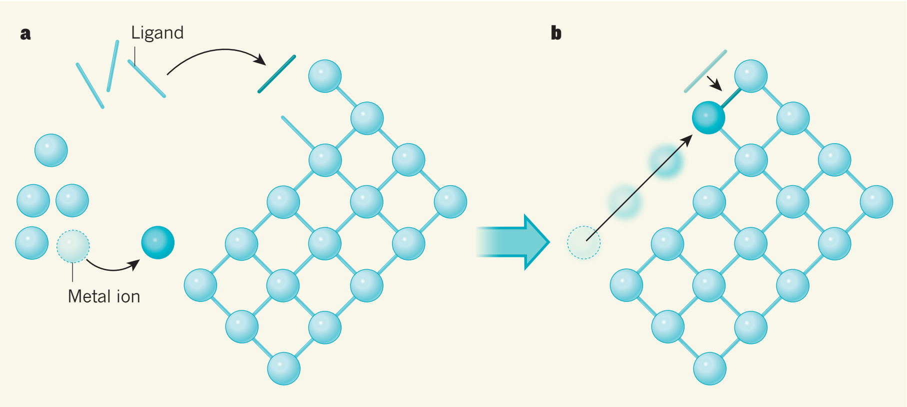


Figure 1 | Growth of a metal–organic framework. Patterson *et al.*[1] have used liquid-cell transmission electron microscopy to study the growth of the ZIF-8 metal–organic framework (MOF) from its constituent metal ions and ligand molecules. They find evidence for a two-step process. **a**, First, the metal ions and ligands diffuse towards a nascent ZIF-8 crystal. **b**, Second, the ions and ligands move to an edge site, where they coordinate with each other, becoming part of the MOF lattice. This is the rate-limiting step of the process. (Figure adapted from ref. 1.)

the crystallization process in real time persists.

Patterson and colleagues' use of liquid-cell transmission electron microscopy (LCTEM) is a big step forward. This technique allows dynamic processes that occur in liquids to be imaged as they happen. It has been used to observe systems such as biological structures[5] and growing nanocrystals[6], but had not previously been applied to MOF syntheses.

Because analytical samples can be damaged by the electron beam used in LCTEM, the authors began by performing a series of control experiments using a zirconium-based MOF called UiO-66, to decouple the effects of beam irradiation on MOF synthesis from the effects of the reaction mechanism. UiO-66 was a good choice for a control because it is easy to synthesize and extremely stable, which meant that it could be prepared ahead of the LCTEM experiments without any risk of it degrading before use. The authors observed that the dissolution or growth of UiO-66 particles depends on the voltage of the electron beam. A threshold dosage of 40,000 electrons per square nanometre was also established — as long as experiments were performed below this limit, damage and particle motion during crystallization were negligible.

Patterson *et al.* then chose another MOF, ZIF-8, as the ideal candidate for demonstrating the LCTEM method. The growth mechanisms of ZIF-8 have previously been studied[7] using transmission electron microscopy on samples removed from synthetic solutions of the MOF, which provided a good comparison with the LCTEM results. ZIF-8 can be synthesized at room temperature in methanol, using zinc nitrate as the metal source and 2-methylimidazole molecules as the organic ligands. The authors observed the growth of ZIF-8 in real time over 11 minutes — the first particles detected were 15 nm in diameter, with subsequent growth observed up to 50 nm.

The authors went on to record videos of the particle formation and used them to determine the growth kinetics of the MOF, uncovering several important features of the crystallization process. First, they proved by direct observation that ZIF-8 particles form through the growth of smaller subunits, rather than by particles coalescing. They also found that an excess of ligand molecules leads to the formation of ZIF-8 particles that are smaller than those formed when the metal-to-ligand ratio is 1:1. The researchers had predicted this using *ex situ* methods, but LCTEM enabled them to observe the process as it occurred.

A series of careful growth experiments was then performed under various accumulative electron doses. The results convincingly show that LCTEM can be applied effectively to study nanoparticles that are easily damaged by electron beams. It remains to be seen whether the technique will be effective at the higher temperatures (typically greater than 100 °C) at which most MOFs form.

A general conclusion from this work is that MOF growth occurs through the transport of metals and ligands to a nascent particle, followed by their movement to an edge or surface site, where bonding between the metal and ligand finally occurs (Fig. 1). The attachment of metal–ligand monomers to a surface site is therefore the controlling factor in particle growth, and the process is not diffusion-limited.

The development of this ability to watch particle formation *in situ* during MOF self-assembly should enable a variety of complicated synthetic questions to be answered. One example is how the addition of 'modulator' compounds, which are sometimes used in MOF syntheses to control the crystallinity of the products, affects the growth kinetics. For MOFs that can adopt different structures, LCTEM could also shed light on what dictates whether kinetic products — those that crystallize most quickly — form during reactions, rather than the most thermodynamically stable products. LCTEM is a much-needed addition to the MOF-characterization toolkit, and its use in conjunction with other methods will no doubt lead to the specific control of crystal morphology, compositions and defects. ■

Krista S. Walton *is at the School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA.*
e-mail: krista.walton@chbe.gatech.edu

1. Patterson, J. P. *et al. J. Am. Chem. Soc.* **137,** 7322–7328 (2015).
2. Zheng, C., Greer, H. F., Chianga, C.-Y. & Zhou, W. *CrystEngComm* **16,** 1064–1070 (2014).
3. Ragon, F., Chevreau, H., Devic, T., Serre, C. & Horcajada, P. *Chemistry* **21,** 7135–7143 (2015).
4. Ahnfeldt, T. *et al. Chemistry* **17,** 6462–6468 (2011).
5. de Jonge, N., Peckys, D. B., Kremers, G. J. & Piston, D. W. *Proc. Natl Acad. Sci. USA* **106,** 2159–2164 (2009).
6. Liao, H.-G., Niu, K. & Zheng, H. *Chem. Commun.* **49,** 11720–11727 (2013).
7. Venna, S. R., Jasinski, J. B. & Carreon, M. A. *J. Am. Chem. Soc.* **132,** 18030–18033 (2010).

MATERIALS SCIENCE

Composite for energy storage takes the heat

A polymer-based material has been discovered that breaks the rules — it has the right combination of properties for use in energy-storage devices called dielectric capacitors, and can function at high temperatures. SEE LETTER P.576

HARRY J. PLOEHN

Devices known as dielectric capacitors have a crucial role in applications that require short, intense power pulses or the conversion of direct current to alternating current. These applications include electronic systems for the integration of energy from renewable sources into power grids[1], transport[2] and military weapon systems[3]. They depend on electrically insulating materials known as dielectrics, which come in several types. Polymeric dielectrics offer advantages for large capacitors, but suffer from low operating temperatures (usually well below 150 °C) and low energy density (which means that devices that use polymeric dielectrics occupy large volumes). On page 576 of this issue, Li *et al.*[4] report that a composite of a polymer and nanometre-scale sheets of boron nitride provides more than a 40% improvement in energy density compared with the best-available polymer dielectric, as well as remarkable stability at temperatures up to 300 °C across a wide range of electric-field frequencies.

Dielectric capacitors achieve the highest rate of energy transfer (termed the power or rate capability) of all capacitor types. They store energy through a variety of molecular and nanoscale electron-polarization mechanisms[5,6] that create oriented dipoles and associated dipolar electric fields. For high energy density, dielectric materials must have a high density of dipoles that have large induced dipole moments (which provide a measure of a charged system's polarity). A dielectric's rate capability depends on how fast charges polarize and depolarize — how fast the dipoles reorient — as an applied electric field varies. Invariably, not all of the energy stored in dipolar electric fields is recovered on depolarization; some is transferred into molecular translation and vibration (thermal energy) and is lost as heat, a process called dielectric loss.

When and how polarized electrons begin to 'leak' (conduct) through a dielectric depends on a property called the dielectric breakdown field strength (E_b). Relatively small leakage currents may occur at field strengths below E_b. Once the field reaches E_b, it promotes a cascade of electrons into the material's conduction band, resulting in catastrophic breakdown as the dielectric is transformed from an insulator into a conductor — which is bad news for dielectric capacitors. Leakage current also converts electrical energy into thermal energy. If this heat is not efficiently removed, the dielectric's internal temperature rises, amplifying molecular motions and degrading the material's mechanical properties to the point that electromechanical stresses create another mechanism for dielectric breakdown, shifting E_b to lower values.

The search for dielectrics that have high energy density, high rate capability and low conversion of electrical energy into heat has followed two distinct, yet intersecting paths. The first has led towards pure, homogeneous materials that can be synthesized and fabricated into large capacitors at minimal cost. The second has led to heterogeneous, multiphase composites, which sacrifice some of the ease of manufacturing simplicity for an optimal compromise between properties and performance. Homogeneous materials include inorganics such as barium titanate (BTO) and organic polymers such as biaxially oriented polypropylene (BOPP, currently the best polymer dielectric) and polyvinylidiene fluoride (PVDF).

Polymers are the preferred choice for large capacitors because of the ease with which they can be processed and their defects controlled. However, the Moss rule, which originated in the semiconductor field[7–9], constrains the dielectric properties of homogeneous materials: an increased polarizability is invariably accompanied by a decreased E_b, which in turn lowers the maximum-attainable volumetric energy density, $\breve{U}$.

Heterogeneous materials might be able to get around the Moss rule. Most studies of such materials are variations on a theme: dispersing a suitable filler material (such as BTO) in a polymer may yield a composite with both high polarizability and high energy density, as well as adequate processability for fabrication into large, reliable capacitors. But often, the addition of the filler dramatically reduces E_b, eliminating any advantage. Dielectric losses and thermal management are also unresolved issues — even though these are key issues in the engineering of large capacitors.

Li and colleagues' work addresses a different set of issues that may enable polymers to bend the Moss rule, if not to break it. Scientists from the same research group previously reported[10] that blends of boron nitride nanosheets (BNNS) with a PVDF-based polymer resulted in remarkable increases in E_b, $\breve{U}$, charge–discharge efficiency (the fraction of stored energy released on discharge), stiffness and thermal conductivity compared with the pure polymer. These improvements were attributed to suppression of leakage currents by the BNNS. Those nanocomposites achieved $\breve{U}$ values up to ten times that of BOPP, but still suffered from high dielectric losses associated with the host polymer.

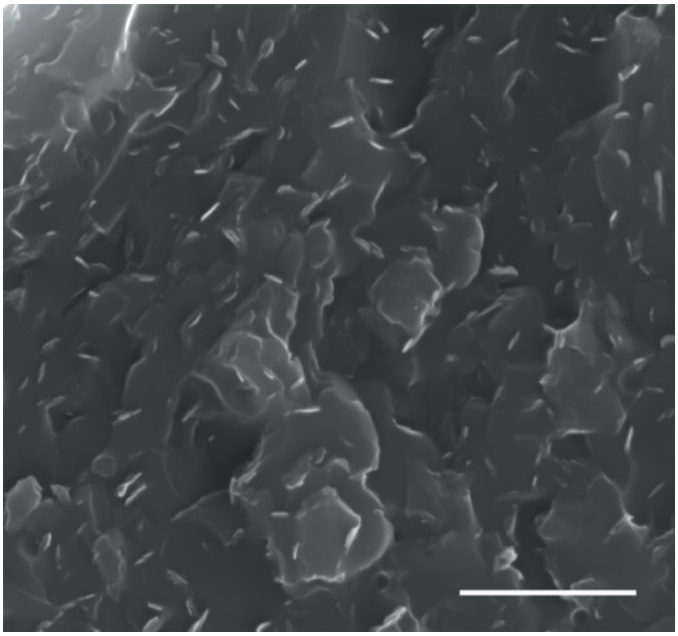


Figure 1 | A high-temperature dielectric material. Li *et al.*[4] have made a composite material in which boron nitride nanosheets (BNNS) are suspended in a polymeric material, cross-linked divinyltetramethyldisiloxane-bis(benzocyclobutene). In this micrograph, the BNNS are visible as lighter flecks against the darker background of the polymer. The composite has better properties as a dielectric material for energy-storage applications than the best-available polymer dielectrics, and operates at higher temperatures. Scale bar, 5 micrometres. (Image from ref. 4.)

In the current work, Li *et al.* blended BNNS with a compound called divinyltetramethyldisiloxane-bis(benzocyclobutene) (BCB), and then reacted the BCB molecules to produce nanocomposites of BNNS in polymeric cross-linked BCB (c-BCB; Fig. 1). The dielectric properties of these composites are remarkably stable over a wide range of temperatures (from room temperature to 300 °C) and field frequencies (100 hertz to 1 MHz). The volumetric energy density and charge–discharge efficiency of c-BCB/BNNS composites at 150 °C greatly exceed those of other polymers designed to work at high temperatures, and maintain meaningful values even at 300 °C — more than 200 °C higher than BOPP's thermal limit for practical use. None of the other candidate high-temperature polymers studied by Li *et al.* approaches this level of dielectric performance at such high temperatures.

Li and co-workers observed that BNNS suppress leakage currents in the nanocomposite by about a factor of ten compared with pristine c-BCB, even at high temperatures. Remarkably, BNNS also increase the polymer's E_b value by 30–50%. More work should be carried out to better understand the underlying mechanisms for suppression of current leakage and dielectric breakdown by BNNS, which form the basis by which c-BCB/BNNS apparently circumvents the Moss rule. Finally, the researchers report that BNNS increase the polymer's thermal conductivity sixfold, and double its stiffness at temperatures above 150 °C. These attributes will help c-BCB/BNNS composites to stay cool and retain electromechanical stability under continuous charge–discharge cycling, reducing or eliminating the need for auxiliary thermal-management systems.

One drawback of the new composites is that BCB will be more expensive to produce than BOPP's monomeric precursor, and blending BNNS into c-BCB represents an additional processing step compared with the manufacture of a pure polymer. It also remains to be seen what effects the BNNS will have on the number of defects and long-term reliability of the composites. Nonetheless, the relatively low filler loading of the nanocomposites, and the fact that irreversible cross-linking can be induced on heating or irradiation with ultraviolet light, will facilitate the development of 'roll-to-roll' processes for producing dielectric polymer films, which may help to control manufacturing costs for capacitors. If the advantages of the BNNS filler can be translated to other dielectric polymers that have higher polarizabilities than c-BCB, then even more remarkable advances in discharged energy density can be anticipated — as well as a reduction in the size of dielectric capacitors for electronics in power systems. ■

Harry J. Ploehn *is in the Department of Chemical Engineering, University of South Carolina, Columbia, South Carolina 29208, USA.*
e-mail: ploehn@mailbox.sc.edu

1. Carrasco J. M. *et al. IEEE Trans. Ind. Electron.* **53,** 1002–1016 (2006).
2. Emadi, A. *et al. IEEE Trans. Power Elect.* **21,** 567–577 (2006).
3. Barshaw, E. J. *et al. IEEE Trans. Magn.* **43,** 223–225 (2010).
4. Li, Q. *et al. Nature* **523,** 576–579 (2015).
5. Raju, G. G. *Dielectrics in Electric Fields* (Dekker, 2003).
6. Nelson, J. K. *Dielectric Polymer Nanocomposites* (Springer, 2010).
7. Ziman, J. M. *Principles of the Theory of Solids* 2nd edn (Cambridge Univ. Press, 1972).
8. Van Vechten, J. A. *Phys. Rev.* **182,** 891–905 (1969).
9. Wemple, S. H. & DiDomenico, M. Jr *Phys. Rev. B* **3,** 1338–1351 (1971).
10. Li, Q. *et al. Energy Environ. Sci.* **8,** 922–931 (2015).

STRUCTURAL BIOLOGY

Arresting developments in receptor signalling

The first crystal structure of a G-protein–coupled receptor in complex with an arrestin protein provides insight into how the signalling pathways activated by these receptors are switched off through desensitization. SEE ARTICLE P.561

JEFFREY L. BENOVIC

G-protein-coupled receptors (GPCRs) play an essential part in mediating signalling in the cells of many organisms. This process is primarily controlled by activation-dependent interactions of GPCRs with three protein families: heterotrimeric guanine-nucleotide-binding proteins (G proteins), GPCR kinases (GRKs) and arrestins. Until now, the only complete structure[1] of a GPCR complex was that of the β_2-adrenergic receptor bound to the G protein Gs, solved in 2011. In this issue, Kang *et al.*[2] (page 561) present the second structure of a GPCR complex, in this case the GPCR rhodopsin bound to arrestin-1.

The binding of a GPCR to a G protein results in the activation and subsequent regulation of downstream effector enzymes that modulate levels of 'second messenger' molecules such as cyclic AMP and calcium. By contrast, the interaction of a GPCR with a GRK promotes phosphorylation of the receptor, which in turn facilitates arrestin binding. This turns off G-protein signalling, a process called desensitization, and promotes cellular internalization of the receptors and arrestin-mediated signalling. The dynamics of these protein–protein interactions are complex and incompletely understood.

The best-characterized GPCR signalling pathway regulates the process of phototransduction in rod cells in the retina of the eye. This pathway involves rhodopsin, the G protein transducin and the effector enzyme cGMP phosphodiesterase (Fig. 1a). Phototransduction is highly regulated by a GRK (GRK1) and an arrestin (arrestin-1); mutations in either of these proteins can lead to a visual defect called Oguchi disease. Rhodopsin was the first GPCR to be crystallized in its basal state[3] as well as in various activated conformations[4–6]. Although some structural insight into the binding of G proteins[5,6] and arrestins[7] to rhodopsin has been gained by using peptides co-crystallized with the receptor, Kang and colleagues' structure of the full protein complex is a significant advance.

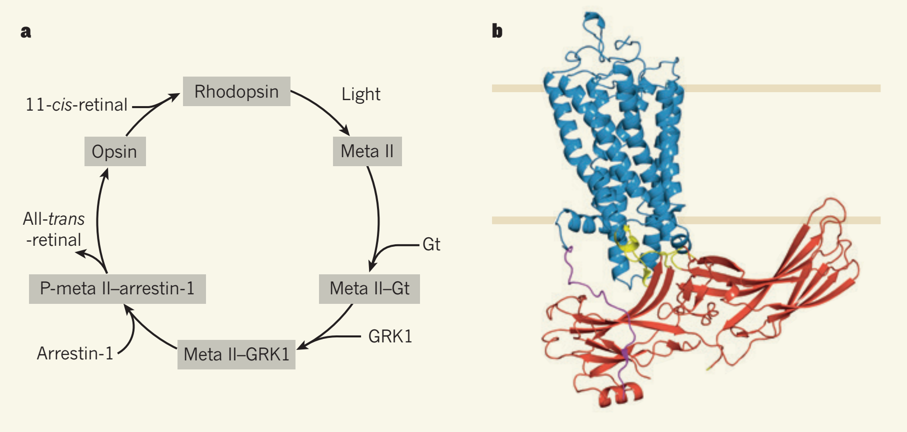


Figure 1 | The phototransduction pathway. **a**, The signalling pathway involving the G-protein-coupled receptor rhodopsin is initiated when light activation induces rhodopsin to form meta II rhodopsin, which interacts with the G protein transducin (Gt) to activate the effector enzyme cGMP phosphodiesterase. This pathway ultimately leads to a rapid visual response in rod cells. Meta II is then phosphorylated by the protein kinase GRK1 to yield P-meta II, which promotes interaction with the protein arrestin-1 to preclude further binding of Gt. P-meta II recycles by losing its bound chromophore (all-*trans*-retinal) to become P-opsin (not shown), which promotes arrestin-1 dissociation and dephosphorylation of the receptor. Opsin then binds the chromophore 11-*cis*-retinal to yield rhodopsin. Each of the individual proteins in this pathway (rhodopsin, meta II, opsin, Gt, GRK1 and arrestin-1) have been crystallized. **b**, Kang *et al.*[2] provide the first crystal structure of a protein complex in this pathway, that of activated rhodopsin (blue) bound to arrestin-1 (red). Sections of arrestin-1 involved in the interface with rhodopsin are shown in yellow. The carboxy terminus of rhodopsin, which is phosphorylated by GRK1, is shown in purple. Rhodopsin is depicted in a phospholipid bilayer.

Solving the X-ray structure of a rhodopsin–arrestin complex proved challenging and required a team of 72 investigators across 25 institutions, who used various tricks to obtain diffractable crystals. First, the two proteins were mutated to aid the formation of active conformations. For rhodopsin, these mutations (E113Q and M257Y) yielded a conformation that was constitutively active even in the absence of a bound chromophore (a colour-determining chemical group, such as all-*trans*-retinal, which typically activates rhodopsin). For arrestin, the researchers mutated three adjacent amino-acid residues in a region that stabilizes the basal conformation and largely overcomes the need for the receptor to be phosphorylated to bind arrestin[8].

The authors were unable to generate a stable rhodopsin–arrestin binary complex (one formed through non-covalent binding). Instead, they purified and crystallized a fusion protein in which arrestin was fused to the carboxy terminus of rhodopsin through a 15-residue linker. This fusion protein also included the enzyme T4 lysozyme at the amino terminus of rhodopsin — this facilitates crystallization without altering the structure of the complex. Finally, because the crystals were small and diffracted to 6–8 ångström in synchrotron experiments, the authors used serial femtosecond X-ray laser crystallography, performed at the SLAC National Accelerator Laboratory in Menlo Park, California, to capture diffraction data for the crystals. This provided enough data to solve a structure of the complex with resolution limits of 3.3–3.8 Å.

The structure reveals multiple points of contact between rhodopsin and arrestin, as well as structural changes in both proteins (Fig. 1b). The primary interface between the proteins involves the finger loop of arrestin (which connects β-strands V and VI in arrestin and adopts an α-helical conformation when bound), which interacts with three regions of rhodopsin: intracellular loop (ICL) 1, the N-terminal region of helix 8, and the C-terminal region of transmembrane (TM) 7. Additional interactions include several arrestin loops: the middle and lariat loops bind to ICL2 on rhodopsin, the back loop binds to TM5, and β-strand VI binds to TM5, TM6 and ICL3.

The authors' basic model proposes that rhodopsin uses multiple structural elements including TM7 and helix 8 to initially recruit arrestin, resulting in a rotation of approximately 20° between the N and C domains of arrestin; this opens a cleft to accommodate rhodopsin's ICL2. Indeed, a similar rotation between the N and C domains has been observed[9] in a preactivated truncated form of arrestin-1 and in β-arrestin-1 (also known as arrestin-2) when bound to a phosphorylated receptor peptide[10]. Kang *et al.* extensively validated their rhodopsin–arrestin structural model by using double electron–electron resonance, hydrogen–deuterium exchange mass spectrometry, cell-based rhodopsin–arrestin interaction assays, and site-specific disulfide cross-linking experiments.

Kang and colleagues' study provides insight into the interactions between GPCRs and arrestins, but much remains to be learned. We clearly need the structures of more GPCR complexes with G proteins, arrestins and GRKs, including complexes of such proteins with the same GPCR — for example, the β_2-adrenergic receptor in complex, separately, with Gs, GRK2 and β-arrestin-1, or rhodopsin in complex with transducin and GRK1 to complement the arrestin-1 structure. Such studies will reveal whether these three classes of protein have specific preferences for particular receptor conformations, and should facilitate the development of compounds that might serve as selective modulators of specific GPCR signalling pathways. This would be an important step in helping to treat the many diseases that are mediated by GPCR signalling pathways, which are currently the target for approximately 40% of the pharmaceuticals on the market. ■

Jeffrey L. Benovic *is in the Department of Biochemistry and Molecular Biology, Thomas Jefferson University, Philadelphia, Pennsylvania 19107, USA.*
e-mail: jeffrey.benovic@jefferson.edu

1. Rasmussen, S. G. F. *et al. Nature* **477,** 549–555 (2011).
2. Kang, Y. *et al. Nature* **523,** 561–567 (2015).
3. Palczewski, K. *et al. Science* **289,** 739–745 (2000).
4. Salom, D. *et al. Proc. Natl Acad. Sci. USA* **103,** 16123–16128 (2006).
5. Scheerer. P. *et al. Nature* **455,** 497–502 (2008).
6. Choe, H.-W. *et al. Nature* **471,** 651–655 (2011).
7. Szczepek, M. *et al. Nature Commun.* **5,** 4801 (2014).
8. Gurevich, V. V. *J. Biol. Chem.* **273,** 15501–15506 (1998).
9. Kim, Y. J. *et al. Nature* **497,** 142–146 (2013).
10. Shukla, A. K. *et al. Nature* **497,** 137–141 (2013).

This article was published online on 22 July 2015.

GENETICS OF DISEASE

Associations with depression

Two genetic regions associated with major depressive disorder have been revealed for the first time, through whole-genome sequencing of a population of Han Chinese women. SEE LETTER P.588

PATRICK F. SULLIVAN

Of all complex human illnesses, major depressive disorder (MDD) has arguably proved the trickiest to understand. Despite decades of research, there is little certainty about its biological basis, in part because genetic clues to its aetiology have been hard to find[1]. The combination of relatively high prevalence and relatively low heritability seems to indicate that MDD does not lend itself to genetic analysis, although the genetic dissection of type 2 diabetes mellitus, which has a similar prevalence and heritability, has been much more productive. On page 588 of this issue, the CONVERGE consortium[2] identifies the first two long-awaited genetic associations for MDD.

Our ignorance about MDD is in marked contrast to its impact on people and public health[3]. The disease is common, costly and associated with high rates of morbidity and mortality. As such, it stands to reason that this research is exciting for those who study MDD. But it also exemplifies a sometimes neglected issue — how an informed approach to improving the definition of a complex illness can lead to success where other approaches have failed.

Why is defining MDD so complicated? Sadness is normal and integral to the human condition. However, much too frequently, sadness becomes pervasive, persistent, unshakable and associated with signs and symptoms characteristic of MDD, such as changes in sleeping habits, appetite and cognition, and the onset of suicidal tendencies. But where should we draw the line between normality and pathology? This question is echoed throughout medicine, for example when using normal fasting blood glucose levels to delineate normal physiology from that of type 2 diabetes, or when separating normal blood pressure from hypertension. The difference is that the measures for these latter two conditions are more objective than those for assessing 'sadness', and the consequences of each disease more readily assessed. There is no laboratory test that will help us to know when sadness becomes MDD.

An informed approach to improving the definition of a complex illness can lead to success where other approaches have failed.

The CONVERGE consortium authors reasoned that the core issue hampering the discovery of MDD-associated genes is heterogeneity — in a group of people who all have the same MDD symptoms, the aetiology of the MDD may in fact be different. Some people might have a highly genetic form of the disease, whereas in others, MDD may be brought on by environmental factors such as poverty, physical or sexual abuse, or an unhealthy lifestyle. Still others may have a primary problem such as alcoholism, of which MDD is a secondary consequence. This long-held concept of heterogeneity has made defining 'true' MDD something of a holy grail.

The researchers made a set of intelligent decisions when defining who to study. They reasoned that more-severe cases would have a clearer and less-complex genetic signal — an approach widely used in human genetics to minimize heterogeneity in complex illnesses.

Unlike more-inclusive approaches[4], the consortium authors implemented several measures that they thought would maximize their chances of success. They worked in China, where the prevalence of MDD is lower than that in the United States or Europe, studied only women and selected relatively severe cases, in which the women had experienced two or more episodes of MDD, using psychiatric inpatient and outpatient facilities. Unusually, they genotyped their samples using low-coverage whole-genome sequencing (genotyping determines the identity of genetic variations across the genome). They identified two regions in which genetic changes, or variants, are associated with MDD — one near the *SIRT1* gene and the other in an intron (a non-protein-coding region) of the gene *LHPP*.

The typical approach to genotyping in human genetics is to use an array containing a fixed set of between 500,000 and 1 million genetic markers — DNA variants at known chromosomal locations. To my knowledge, this is the only published study in which genotyping involved low-coverage sequencing of the whole genome. Because of decreases in the costs of genotyping arrays, it may be one of the last. The authors' low-coverage sequencing had relatively high error rates — around 2% of the genetic variants that they identified could not be replicated with a different method, compared with less than 0.5% for an inexpensive array. Moreover, despite their wish to gain traction on MDD-causing genetic associations that have yet to be described in China, the two variants that they found have been in standard databases for a decade. This is because genetic variants such as these, which are common in China, are evolutionarily old, and so likely to be found across the globe. Wisely, the authors confirmed the variants with a second method and replicated the results in an independent sample, ensuring that their associations meet typical standards for significance and replication.

The consortium suggests that the proximity of one of the variants to *SIRT1* implicates abnormalities in mitochondria (the cell's energy-producing centres) in MDD, because one role of the SIRT1 protein is to regulate mitochondrial function. If this assertion holds up, it is likely to imply that many other genetic variants are involved in altered mitochondrial function and have associations with MDD that near the threshold for significance. A standard way to evaluate such a hypothesis is to investigate whether a specific genetic pathway, such as that involving the genes that affect mitochondrial function, is statistically more likely to have smaller *P*-values for association with MDD than expected by chance. This analysis was not reported in the current study. A previous systematic pathway analysis of MDD and other major psychiatric disorders did not implicate mitochondrial biology[5]. As such, although the researchers' hypothesis is intriguing, it requires replication, extension, integrated analysis and more biological evidence.

I wish the authors had formally tested their fundamental premise, namely that their sample was more homogeneous than those studied previously. If that is the case, then the heritability of the common variants (the proportion of variance contributing to MDD that can be accounted for by the genetic variation they measured) should be notably high. But although several methods exist to check this[6,7], the authors did not report such an analysis.

More unsettling, the two variants identified have almost no association signal in samples taken from European populations[8]. The reasons for this discrepancy are unclear. Perhaps these variants are truly causative for MDD only in severe cases from China. However, many other common associations for complex illnesses hold across the world. The authors tested the comparability of their findings with European samples and found some evidence of overlap, but more-refined analyses would be of keen interest.

This first identification of replicable, significant genome-wide associations for MDD is exceptional. Although further work is required, it is to be hoped that these results will provide therapeutic targets for MDD. The drug-discovery pipeline for MDD has never been based on solid biological foundations, but the work begun here could improve the focus of the field.

The authors' study marks the beginning of the beginning for the genetic dissection of MDD. The CONVERGE consortium has provided an excellent starting point for what should be an intriguing voyage of discovery. ■ See go.nature.com/lsghoc for a related News story.

Patrick F. Sullivan *is in the Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, North Carolina 27599-7264, USA, and in the Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.*
e-mail: pfsulliv@med.unc.edu

1. Levinson, D. F. *et al. Biol. Psychiatry* **76,** 510–512 (2014).
2. CONVERGE consortium. *Nature* **523,** 588–591 (2015).
3. Whiteford, H. A. *et al. Lancet* **382,** 1575–1586 (2013).
4. Kendler, K. S. *Arch. Gen. Psychiatry* **54,** 299–304 (1997).
5. The Network and Pathway Analysis Subgroup of the Psychiatric Genomics Consortium. *Nature Neurosci.* **18,** 199–209 (2015).
6. Cross-Disorder Group of the Psychiatric Genomics Consortium. *Nature Genet.* **45,** 984–994 (2013).
7. Bulik-Sullivan, B. K. *et al. Nature Genet.* **47,** 291–295 (2015).
8. Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium. *Mol. Psychiatry* **18,** 497–511 (2013).

This article was published online on 15 July 2015.

OPHTHALMOLOGY

Cataracts dissolved

Mutations underlying hereditary cataracts in two families impair the function of an enzyme that synthesizes the lens molecule lanosterol. The finding may lead to non-surgical prevention and treatment of cataracts. SEE LETTER P.607

J. FIELDING HEJTMANCIK

In this issue, Zhao *et al.*[1] (page 607) identify a mutation in the gene encoding the enzyme lanosterol synthase (LSS) as the cause of inherited cataracts in two families. LSS, which is produced in the lens, synthesizes lanosterol, a molecule that is amphipathic (that is, it has both hydrophilic and hydrophobic properties). The authors show that lanosterol can dissolve the precipitates, and even the amyloid-like fibril structures, of mutant lens crystallin proteins that are the cause of cataracts in some individuals. Furthermore, lanosterol effectively treated naturally occurring cataracts in rabbit lenses and in dogs *in vivo*. In addition to elucidating the visual process, this work promises to continue in the tradition of lens research by expanding scientific insight into broad and often seemingly unrelated areas of enquiry.

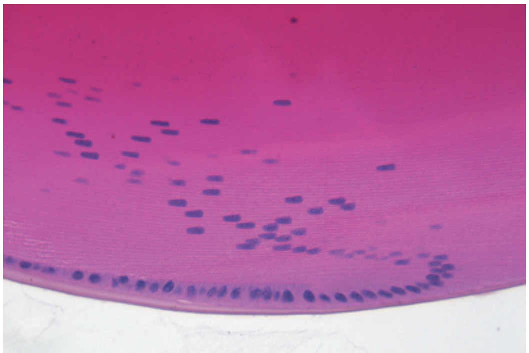
The eye lens has been intensively studied for almost two centuries. In 1833, optics scientist David Brewster deduced the fine structure of the cod lens, calculating that it contained 5 million fibre cells, each 4.8 millimetres long, using only a candle and a finely ruled steel bar[2]. In 1901, embryologist Hans Spemann's study of lens development resulted in the concept of inductive cellular interactions during embryonic development[3]. Studies of lens biochemistry began in the late nineteenth century with descriptions of the high concentrations of heterogeneous structural proteins now known as crystallins[4]. Subsequently, one of the first genetic locations on a non-sex chromosome to be associated with disease was linked to cataract susceptibility[5], and messenger RNA molecules encoding chicken lens δ-crystallins were among the first mRNAs to be isolated, cloned and studied[6].

The function of the eye lens is to transmit light and focus it on the retina. The lens accomplishes this through a single cell type that follows a developmental pattern, beginning as a member of the germinative zone in the single layer of anterior epithelial cells overlaying a mass of fibre cells. The epithelial cells then migrate laterally towards the lens equator, where they elongate and invert to form secondary fibre cells, arranged in a curved, onion-like configuration. As they do this, the cells synthesize large amounts of crystallins, such that they contain perhaps the highest concentration of proteins found in any tissue. They also degrade organelles, minimize extracellular space and increase the density of their cell membranes to levels approaching that of the cell's cytoplasm, all of which decrease light scattering[7]. Thus, transparency is accomplished largely through a combination of the microarchitecture of the lens and, on a molecular level, the densely packed lens crystallins (Fig. 1).

Human crystallins are divided into two families, α- and βγ-crystallins; together, these make up 90% of the water-soluble proteins in lens cells[8]. They are extremely stable, highly ordered and provide a relatively constant refractive index, which allows lens transparency[9]. Because differentiated lens fibre cells lack the synthetic apparatus to produce new proteins, crystallins are not turned over, and those in the centre of the lens are among the oldest proteins in the body. Preserving crystallin structure and function is therefore crucial for prevention of lens opacities. Other biological activities of the lens serve primarily to protect the complementary systems of crystallin packing and fibre-cell arrangement from disruption and damage by age and external insults, especially ultraviolet light, oxidative stress and glycation.

The genes that cause cataracts when mutated tend to encode proteins that are involved in one of these biological pathways or functional groupings of proteins that are critical for lens homeostasis. In families at risk of congenital cataracts for which the mutant gene is known, slightly less than half have mutations in lens crystallins, with others having mutations in growth or transcription factors, membrane proteins, chaperone proteins and proteins involved in protein degradation, among others[10]. Zhao and colleagues' identification of LSS mutations as a cause of congenital cataracts adds a new pathway.

The catastrophic structural changes in crystallins seen in many hereditary cataracts can overwhelm the defensive systems of the lens, and might also be refractory to the solubilizing activity of lanosterol identified by the authors. Nevertheless, this agent might be more therapeutically applicable to the slow progressive denaturation of crystallins seen in age-related cataracts. In age-related cataracts, damaged βγ-crystallin proteins are bound by α-crystallins, which act like chaperones — proteins that assist the folding or unfolding of other proteins — except that, instead of refolding the denatured βγ-crystallins, α-crystallins solubilize them[11], thereby reducing light scattering. However, as more crystallins are damaged and bound over time, the protein complexes themselves become large enough to scatter light[11,12]. Eventually, the complexes precipitate, forming the insoluble protein fraction (termed high molecular weight aggregates) that increases with normal ageing and especially in cataractous lenses. This identifies cataracts, in at least some cases, as a protein-misfolding disease[13].



PAUL FITZGERALD/REF. 16

Figure 1 | The eye lens. This cross-section of a mouse eye lens shows the curved, onion-like configuration of fibre cells, which are packed close together and lose subcellular structures such as nuclei (coloured blue) as they mature and move to the centre of the lens. Fibre cells contain highly ordered crystallin proteins, the intracellular concentration of which increases towards the interior of the lens (seen as darkening pink). This combined cellular and intracellular structure gives transparency to the lens. Denaturation and aggregation of crystallin proteins can result in the lens opacity known as a cataract. Zhao *et al.*[1] show that the molecule lanosterol can redissolve crystallin aggregates and alleviate cataracts.

Although surgery to remove cataracts is efficacious and safe, ageing populations around the world are predicted to require a doubling of cataract surgery in the next 20 years[14]. The same population demographics suggest that, if development of age-related cataracts in susceptible individuals could be delayed by even ten years, the need for surgery could be reduced by almost half[15]. Pre-symptomatic screening of age-related cataracts is easy, and the eye is easily accessible for topical application of drugs. Zhao and colleagues show that eye drops containing lanosterol successfully treated natural cataracts in dogs. The potential for this finding to be translated into the first practical pharmacological prevention, or even treatment, of human cataracts could not come at a more opportune time. Furthermore, this approach might serve as a model for other protein-misfolding diseases affecting a variety of tissues and organ systems. ■

J. Fielding Hejtmancik *is in the Ophthalmic Molecular Genetics Section, Ophthalmic Genetics and Visual Function Branch, National Eye Institute, Rockville, Maryland 20892-9402, USA.*
e-mail: hejtmancikj@nei.nih.gov

1. Zhao, L. *et al. Nature* **523,** 607–611 (2015).
2. Brewster, D. *Phil. Trans. R. Soc. Lond.* **123,** 323–332 (1833).
3. Spemann, H. *Vehr. Anat. Ges.* **15,** 61–79 (1901).
4. Zhang, T. *et al. Hum. Mutat.* **30,** E603–E611 (2009).
5. Renwick, J. H. & Lawler, S. D. *Ann. Hum. Genet.* **27,** 67–84 (1963).
6. Zelenka, P. S. & Piatigorsky, J. *Proc. Natl Acad. Sci. USA* **71,** 1896–1900 (1974).
7. Michael, R., van Marle, J., Vrensen, G. F. & van den Berg, T. J. *Exp. Eye Res.* **77,** 93–99 (2003).
8 Bloemendal, H. *et al. Prog. Biophys. Mol. Biol.* **86,** 407–485 (2004).
9. Benedek, G. B. *Appl. Optics* **10,** 459–473 (1971).
10 Shiels, A. & Hejtmancik, J. F. *Clin. Genet.* **84,** 120–127 (2013).
11. Rao, P. V., Huang, Q.-L., Horwitz, J. & Zigler, J. S. Jr *Biochim. Biophys. Acta* **1245,** 439–447 (1995).
12. Datiles, M. B. III *et al. Arch. Ophthalmol.* **126,** 1687–1693 (2008).
13. Moreau, K. L. & King, J. A. *Trends Mol. Med.* **18,** 273–282 (2012).
14. Taylor, H. R. *Br. J. Ophthalmol.* **84,** 1–2 (2000).
15. Kupfer, C. *Invest. Ophthalmol. Vis. Sci.* **28,** 2–8 (1987).
16. Sun, N., Shibata, B., Hess, J. F. & FitzGerald, P. G. *Mol. Vis.* **21,** 428–442 (2015).

This article was published online on 22 July 2015.

STRONG-FIELD PHYSICS

Harmonic radiation from crystals

Electrons in a crystal can tunnel between energy bands when a strong electric field is switched on. It emerges that electron pathways interfere almost instantaneously, giving rise to ultra-short, pulsed emission of light. See Letter p.572

PETER HOMMELHOFF & TAKUYA HIGUCHI

The puzzling but experimentally verified fact that particles can propagate through walls is a hallmark of quantum mechanics. In crystalline solids, the motion of electrons is restricted by the presence of the atomic lattice, which limits their energy to certain ranges known as energy bands. Because an electron's energy cannot exceed these limits, gaps are formed between bands.

In the heyday of quantum theory, the physicist Clarence Zener showed[1] that the electrons in a solid that is subjected to a strong electric field can tunnel between energy bands, traversing the classically imposed barrier. On page 572 of this issue, Hohenleutner *et al.*[2] report an experimental and theoretical study showing that, when a strong electric field oscillating at terahertz frequencies (1 THz is 10^{12} Hz) is applied to a crystal, various bands can be coupled together by electron tunnelling, and the crystal emits ultra-short bursts of high-harmonic radiation.

The frequency of this oscillating driving field, which corresponds to mid-infrared wavelengths, is much lower than any of the frequencies required for straightforward electron jumps between different energy bands. Because of its large strength, however (greater than 1 volt per nanometre), the field drives electrons to tunnel from one band to another[3,4] on femtosecond timescales (1 fs is 10^{-15} seconds), that is, almost instantaneously with the switching-on of the field. The electrons' dynamics are complex and include not only tunnelling to different energy bands, but also acceleration within each band; these processes result in the radiation of electromagnetic waves at a much higher frequency than that of the driving terahertz-frequency field. The emitted radiation is called high-harmonic radiation because its spectrum usually displays peaks at harmonics (integer multiples) of the driving field's frequency, reflecting the field's temporal periodicity.

The observed high-harmonic radiation consists of pulses of ultra-broadband visible

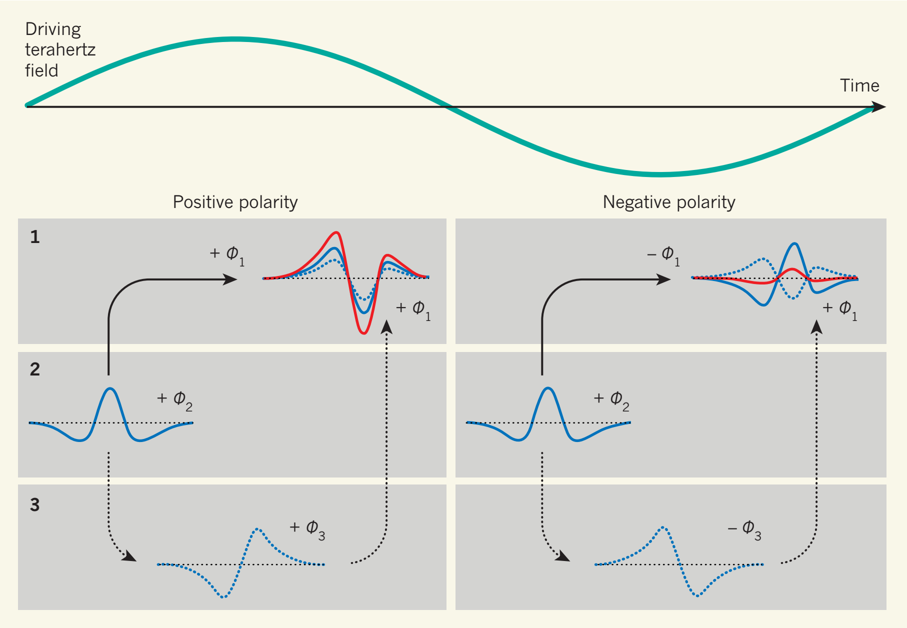


Figure 1 | Electron tunnelling and interference. The electrons in a solid form energy bands (here, 1–3) and can tunnel between these when a strong, terahertz-frequency field is applied. Band 1 can be reached directly from band 2 (solid arrow), but also through the path 2 to 3 to 1 (dotted arrows). The sign (+ or −) of the electrons' wavefunctions (Φ; solid and dotted blue lines) after each tunnelling event is determined by the polarity of the applied field (one cycle, shown in green). Hohenleutner *et al.*[2] observe that the electrons' wavefunction does not change sign after each tunnelling event for positive field polarity (left), but does change sign for negative polarity (right). As a result, for positive polarity the wavefunctions interfere constructively (red line in band 1, left), whereas for negative polarity they interfere destructively (red line in band 1, right). Consequently, for positive polarities of the applied field only, the crystal emits a pulse of radiation (not shown) at frequencies corresponding to high harmonics of the applied field's frequency.

and infrared light that are only several femtoseconds long. These pulses reveal the signature of the electronic states populated by the tunnelling process and allow accurate tracing of the dynamics of the crystal's electrons. The authors record the high-harmonic radiation that is emitted from a gallium selenide crystal subjected to pulses centred at a frequency of about 30 THz. Their experiment measures the structure of the high-harmonic spectrum with femtosecond precision, providing greater insight into the electron dynamics than would be possible by measurements of the spectrum without ultrafast temporal resolution. Using sophisticated optical techniques, the authors are able to pinpoint the moment at which the high harmonics are generated during a terahertz-frequency pulse that is only a few cycles long.

These results would have been hard to interpret without the in-depth theoretical understanding and modelling that Hohenleutner and colleagues use to complement their experiment. The data can be explained by invoking inter-band tunnelling, but the authors show that more than two electron energy bands are involved — five are required. Moreover, quantum interference between the various tunnelling paths needs to be invoked for a proper explanation of the observations. In classic interference, waves combine constructively or destructively depending on whether they arrive at a particular spot in or out of phase, which hinges on the difference between the distances the waves have covered. Analogously, in this experiment, different electron-excitation pathways give rise to quantum interference and couple the five energy bands together.

Using numerical modelling, the authors artificially 'switched off' the interference between the tunnelling pathways, demonstrating that the quantum interference is essential to fit the experimental data. The modelling also shows that the polarity of the driving electric field determines whether the interference is constructive or destructive (Fig. 1). It turns out that high-harmonic radiation is emitted only when the driving electric field is at its peak, and for one polarity of the field only. When this happens, the intensity of the emitted high-harmonic pulse is enhanced by a factor of 30, owing to constructive interference between the tunnelling pathways, compared with the case without interference.

High-harmonic emission in solids has been a vibrant area of research since it was shown that the excitation of crystals by a strong, long-wavelength electromagnetic field generates radiation at high-harmonic frequencies of the driving field's fundamental frequency[5,6]. The strong-field physics of individual atoms in the gas phase is well understood, but the different scaling of the maximum harmonic frequency with the strength of the incident field in solids had hinted at different underlying processes[5]. However, a recent comparative study[7] found that some aspects of the gas-physics picture apply to solids too, such as the importance of controlling the electrons' pathways so as to generate high-harmonic radiation. In another study[8], researchers applied optical pulses of sub-period duration to a silicon dioxide crystal and showed that the observed high-harmonic radiation reaches frequencies of 8,500 THz (corresponding to the extreme- ultraviolet domain), a record value for intra-band currents induced in a solid.

From insights into such observations, new ways may be devised to control the phase of the electron wavefunction in crystals**,** for instance through the instantaneous modification of the band structure by external electric fields. Related work with solids subjected to strong fields has also shown that the material may be reversibly changed from being a dielectric (insulator) to a semi-metal (conductor) in a femtosecond[9–11].

These studies, and Hohenleutner and colleagues' work in particular, open the door to using electrons in solids as a quantum-physics playground. For example, fully understanding the radiation mechanism will allow us to infer the electron wavefunctions from the emitted high harmonics. This might enable the tomographic reconstruction of a crystal's electronic band structure. Moreover, the ultra-short timescales observed in this work could open up new ways of quantum information processing, in which information will be encoded in the electrons' wavefunctions. Other applications can also be envisaged, including the generation of intense sources of coherent extreme-ultraviolet radiation. Because of the complex nature of any condensed material, further strong-field studies of crystals, disordered solids, and even liquids might lead to other surprise discoveries. ■

Peter Hommelhoff *and* **Takuya Higuchi** *are in the Department of Physics, Friedrich–Alexander–Universität (FAU) Erlangen–Nürnberg, Erlangen 91058, Germany.*
e-mail: peter.hommelhoff@fau.de

1. Zener, C. *Proc. R. Soc. Lond. A* **145,** 523–529 (1934).
2. Hohenleutner, M. *et al. Nature* **523,** 572–575 (2015).
3. Keldysh, L. V. *Sov. Phys. JETP* **20,** 1307–1314 (1965).
4. Ghimire, S. *et al. J. Phys. B.* **47,** 204030 (2014).
5. Ghimire, S. *et al. Nature Phys.* **7,** 138–141 (2011).
6. Schubert, O. *et al. Nature Photon.* **8,** 119–123 (2014).
7. Vampa, G. *et al. Nature* **522,** 462–464 (2015).
8. Luu, T. T. *et al. Nature* **521,** 498–502 (2015).
9. Durach, M., Rusina, A., Kling, M. F. & Stockman, M. I. *Phys. Rev. Lett.* **107,** 086602 (2011).
10. Schiffrin, A. *et al. Nature* **493,** 70–74 (2013).
11. Schultze, M. *et al. Science* **346,** 1348–1352 (2014).

CORRECTION

The News & Views article 'Astrochemistry: Fullerene solves an interstellar puzzle' by Pascale Ehrenfreund and Bernard Foing (*Nature* **523,** 296–297; 2015) omitted the relevant reference citations and full credit information for Figure 1. This has now been corrected online.

doi:10.1038/nature14565

Timing and climate forcing of volcanic eruptions for the past 2,500 years

M. Sigl[1]†, M. Winstrup[2], J. R. McConnell[1], K. C. Welten[3], G. Plunkett[4], F. Ludlow[5], U. Büntgen[6,7,8], M. Caffee[9,10], N. Chellman[1], D. Dahl-Jensen[11], H. Fischer[7,12], S. Kipfstuhl[13], C. Kostick[14], O. J. Maselli[1], F. Mekhaldi[15], R. Mulvaney[16], R. Muscheler[15], D. R. Pasteris[1], J. R. Pilcher[4], M. Salzer[17], S. Schüpbach[7,12], J. P. Steffensen[11], B. M. Vinther[11] & T. E. Woodruff[9]

Volcanic eruptions contribute to climate variability, but quantifying these contributions has been limited by inconsistencies in the timing of atmospheric volcanic aerosol loading determined from ice cores and subsequent cooling from climate proxies such as tree rings. Here we resolve these inconsistencies and show that large eruptions in the tropics and high latitudes were primary drivers of interannual-to-decadal temperature variability in the Northern Hemisphere during the past 2,500 years. Our results are based on new records of atmospheric aerosol loading developed from high-resolution, multi-parameter measurements from an array of Greenland and Antarctic ice cores as well as distinctive age markers to constrain chronologies. Overall, cooling was proportional to the magnitude of volcanic forcing and persisted for up to ten years after some of the largest eruptive episodes. Our revised timescale more firmly implicates volcanic eruptions as catalysts in the major sixth-century pandemics, famines, and socioeconomic disruptions in Eurasia and Mesoamerica while allowing multi-millennium quantification of climate response to volcanic forcing.

Volcanic eruptions are primary drivers of natural climate variability—their sulfate aerosol injections into the stratosphere shield the Earth's surface from incoming solar radiation, leading to short-term cooling at regional-to-global scales[1]. Temperatures during the past 2,000 years have been reconstructed at regional[2], continental[3], and global scales[4] using proxy information from natural archives. Tree-ring-based proxies provide the vast majority of climate records from mid- to high-latitude regions of (predominantly) the Northern Hemisphere, whereas ice-core records (for example, $\delta^{18}O$) represent both polar regions[3].

Climate forcing reconstructions for the Common Era (CE)—including solar (for example, ^{10}Be)[5] and volcanic (for example, sulfate)[6,7] activity—derive mostly from ice-core proxies. Any attempt to attribute reconstructed climate variability to external volcanic forcing, and to distinguish between response, feedback, and internal variability of the climate system, requires ice-core chronologies that are synchronous with those of other climate reconstructions. In addition, multi-proxy climate reconstructions[2–4] derived from ice cores and other proxies such as tree rings will have diminished high- to mid-frequency amplitudes if the individual climate records are on different timescales.

The magnitudes and relative timing of simulated Northern Hemisphere temperature responses to large volcanic eruptions are in disagreement with reconstructed temperatures obtained from tree rings[8,9], but it is unclear to what extent this model/data mismatch is caused by limitations in temperature reconstructions, volcanic reconstructions, or implementation of aerosol forcing in climate models[9–11]. The hypothesis of chronological errors in tree-ring-based temperature reconstructions[8,9] offered to explain this model/data mismatch has been tested and widely rejected[11–14], while new ice-core records have become available providing different eruption ages[15,16] and more precise estimates of atmospheric aerosol mass loading[17] than for previous volcanic reconstructions.

Using documented[18] and previous ice-core-based eruption ages[16], strong summer cooling following large volcanic eruptions has been recorded in tree-ring-based temperature reconstructions during the second millennium CE with a one-to-two year lag similar to that observed in instrumental records after the 1991 Pinatubo eruption[19]. An apparent seven-year delayed cooling observed in individual tree-ring series relative to Greenland ice-core acidity peaks during the first millennium CE, however, suggests a bias in existing ice-core chronologies[20,21]. Using published ice-core chronologies, we also observed a seven-year offset between sulfate deposition in North Greenland and the start of tree-ring growth reduction in a composite of five multi-centennial tree-ring summer temperature reconstructions ('N-Tree') from the Northern Hemisphere between 1 and 1000 CE (Methods), whereas tree-ring response was effectively immediate for eruptions occurring after 1250 CE (Fig. 1a).

Precise time marker across hemispheres

Independent age markers with which to test the accuracy of tree-ring and ice-core chronologies have recently become available with the detection of abrupt enrichment events in the ^{14}C content of tree rings. Rapid increases of atmospheric ^{14}C were first identified in individual growth increments of cedars from Japan between 774 CE and 775 CE[22] and between 993 CE and 994 CE[23]. The presence and timing of

[1]Desert Research Institute, Nevada System of Higher Education, Reno, Nevada 89512, USA. [2]Department of Earth and Space Sciences, University of Washington, Seattle, Washington 98195, USA. [3]Space Sciences Laboratory, University of California, Berkeley, California 94720, USA. [4]School of Geography, Archaeology and Palaeoecology, Queen's University Belfast, Belfast BT7 1NN, UK. [5]Yale Climate and Energy Institute, and Department of History, Yale University, New Haven, Connecticut 06511, USA. [6]Swiss Federal Research Institute WSL, 8903 Birmensdorf, Switzerland. [7]Oeschger Centre for Climate Change Research, University of Bern, 3012 Bern, Switzerland. [8]Global Change Research Centre AS CR, 60300 Brno, Czech Republic. [9]Department of Physics, Purdue University, West Lafayette, Indiana 47907, USA. [10]Department of Earth, Atmospheric, and Planetary Sciences, Purdue University, West Lafayette, Indiana 47907, USA. [11]Centre for Ice and Climate, Niels Bohr Institute, University of Copenhagen, 2100 Copenhagen, Denmark. [12]Climate and Environmental Physics, University of Bern, 3012 Bern, Switzerland. [13]Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, 27570 Bremerhaven, Germany. [14]Department of History, The University of Nottingham, Nottingham NG7 2RD, UK. [15]Department of Geology, Quaternary Sciences, Lund University, 22362 Lund, Sweden. [16]British Antarctic Survey, Natural Environment Research Council, Cambridge CB3 0ET, UK. [17]The Laboratory of Tree-Ring Research, University of Arizona, Tucson, Arizona 85721, USA. †Present address: Laboratory of Radiochemistry and Environmental Chemistry, Paul Scherrer Institut, 5232 Villigen, Switzerland.

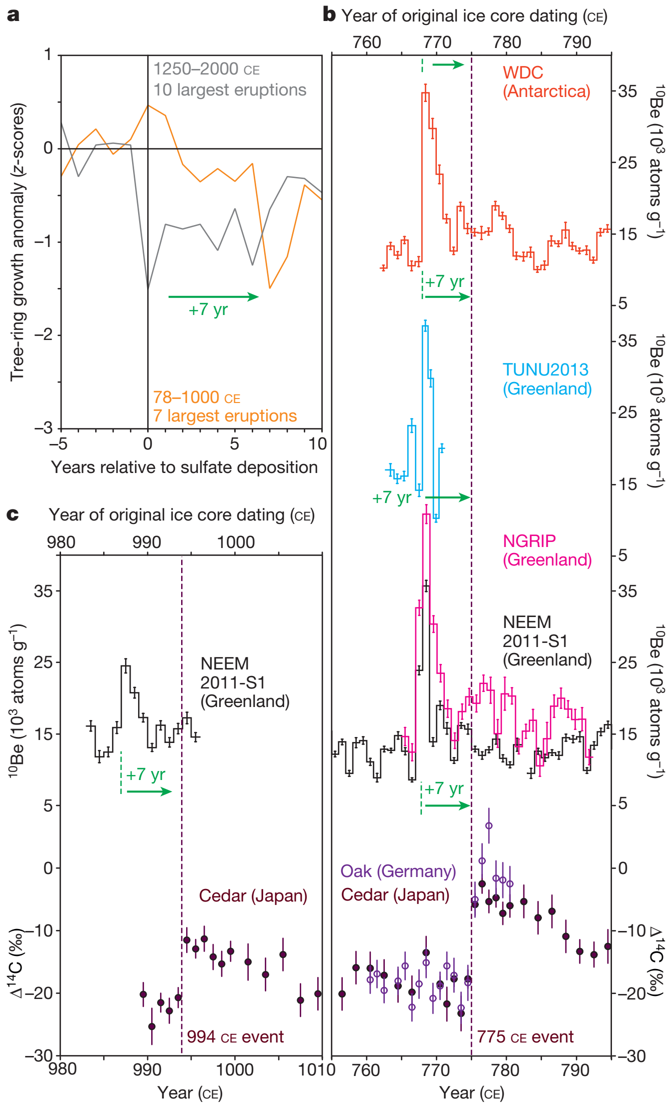


Figure 1 | Annual ^{10}Be ice-core records and post-volcanic cooling from tree rings under existing ice-core chronologies. **a**, Superposed epoch analysis for the largest volcanic signals in NEEM-2011-S1 between 78 and 1000 CE ($n = 7$; orange trace) and for the largest eruptions between 1250 and 2000 CE ($n = 10$; grey trace)[16]. Shown are standardized growth anomalies (z scores relative to 1000–1099 CE) from a multi-centennial, temperature-sensitive tree-ring composite (N-Tree[42,43,76–78], Methods) ten years after the year of volcanic sulfate deposition at the NEEM ice core site in Greenland (GICC05 timescale), relative to the level five years before sulfate deposition. **b**, Annually resolved ^{10}Be concentration records from the WDC, TUNU2013, NGRIP, and NEEM-2011-S1 ice cores on their original timescales and annually resolved Δ^{14}C series from tree-ring records between 755 CE and 795 CE[22,24], with green arrows representing the suggested time shifts for synchronization; error bars are 1σ measurement uncertainties; the estimated relative age uncertainty for TUNU2013 at this depth interval from volcanic synchronization with NEEM-2011-S1 is ±1 year. **c**, Annually resolved ^{10}Be concentration record from NEEM-2011-S1 ice core on its original timescale and annually resolved Δ^{14}C series from tree rings in 980 CE and 1010 CE[23]; error bars are 1σ measurement uncertainties.

the event in 775 CE (henceforth, the 775 event) has been reproduced by all radiocarbon measurements performed on tree rings at annual (or higher) resolution—including tree cores from Germany[24], the Alps[12], the Great Basin[25] (USA), and Siberia[25]. Recent identification of the same 775 CE event in kauri wood samples from New Zealand in the Southern Hemisphere demonstrates the global extent of the rapid ^{14}C increase and provides further constraints on the event's exact timing (March 775 ± 6 months) owing to the asynchronous Southern Hemisphere growing season[26]. While the cause of the 775 and 994 events is still debated[22,24,27], we expect that they might also have produced an excess of cosmogenic ^{10}Be through the interaction of high-energy particles with atmospheric constituents[28,29]. Since both of these radionuclides are incorporated rapidly into proxy archives via aerosol deposition (^{10}Be in ice cores) and photosynthesis ($^{14}CO_2$ in tree rings), isotopic anomalies caused by these extraterrestrial events provide a global age marker with which to link ice-core records to tree-ring chronologies directly[27]. The latter serve as an absolute and precise age marker, verified (at least since 775 CE) by the coherence of the rapid increase in ^{14}C in all tree-ring records for which high-resolution radiocarbon analyses were performed, including those speculated to be at risk of missing rings[8].

We measured ^{10}Be concentrations at approximately annual resolution in four ice cores—NEEM-2011-S1, TUNU2013, and NGRIP in Greenland, and the West Antarctic Ice Sheet Divide Core (WDC)—over depth ranges encompassing the year 775 CE as dated in existing ice-core chronologies in order to provide a direct, physically based test of any dating bias in these chronologies (Fig. 1, Extended Data Fig. 1, Methods, Supplementary Data 1). Both polar ice sheets contain ^{10}Be concentrations exceeding the natural background concentration (>150%; 6σ) for one-to-two consecutive years, compatible with the 775 CE event observed in tree rings. Using the original ice-core age models[16,30], the ages of the ^{10}Be maxima in NEEM-2011-S1, NGRIP, and WDC are 768 CE, offset by 7 years from the tree-ring event. A further ^{10}Be anomaly measured in NEEM-2011-S1 at 987 CE, compatible with the 994 CE event in tree rings, suggests that a chronological offset was present by the end of the first millennium CE (Fig. 1). Several different causes may have contributed to the offset (see a summary in the Methods section), among which is the use of a previous dating constraint[30] for Greenland, where volcanic fallout in the ice was believed to indicate the historic (79 CE) eruption of Vesuvius.

Revised ice-core chronologies

Given the detection of a bias in existing ice-core chronologies, we developed new timescales before the 1257 Samalas eruption in Indonesia[31] using highly resolved, multi-parameter aerosol concentration records from three ice cores: NEEM-2011-S1, NEEM, and WDC. We used the StratiCounter program, an automated, objective, annual-layer detection method based on Hidden Markov Model (HMM) algorithms[32] (Methods). For NEEM-2011-S1, the confidence intervals obtained for the layer counts allowed us to improve the dating further by constraining the timescale using the 775 CE ^{10}Be anomaly and three precisely dated observations of post-volcanic aerosol loading of the atmosphere (Fig. 2, Extended Data Tables 1–3, Methods, Supplementary Data S2).

We evaluated the accuracy of our new chronologies ('WD2014' for WDC and 'NS1-2011' for NEEM) by comparison to (1) an extensive database of historical volcanic dust veil observations (Extended Data Fig. 2, Methods, Supplementary Data 2), (2) ice-core tephra evidence (Methods), and (3) the 994 CE event (Methods, Fig. 2). Using the new timescales, we found large sulfate signals in Greenland (for example, in 682 CE, 574 CE, and 540 CE) between 500 CE and 2000 CE that frequently occurred within one year of comparable—and independently dated—signals in Antarctica. These bipolar signals can now be confidently attributed to large tropical eruptions (Fig. 2). Back to 400 BCE, other large sulfate peaks (for example, 44 BCE) were synchronous to within three years (Fig. 2). We conclude that the revised ice-core timescales are accurate to within less than five years during the past 2,500 years, on the basis of combined evidence from radionuclides, tree rings, tephra analyses, and historical accounts. Compared to the previous chronologies, age models differ before 1250 CE by up to 11 years (GICC05, Greenland) and 14 years (WDC06A-7, Antarctica) (Extended Data Fig. 3).

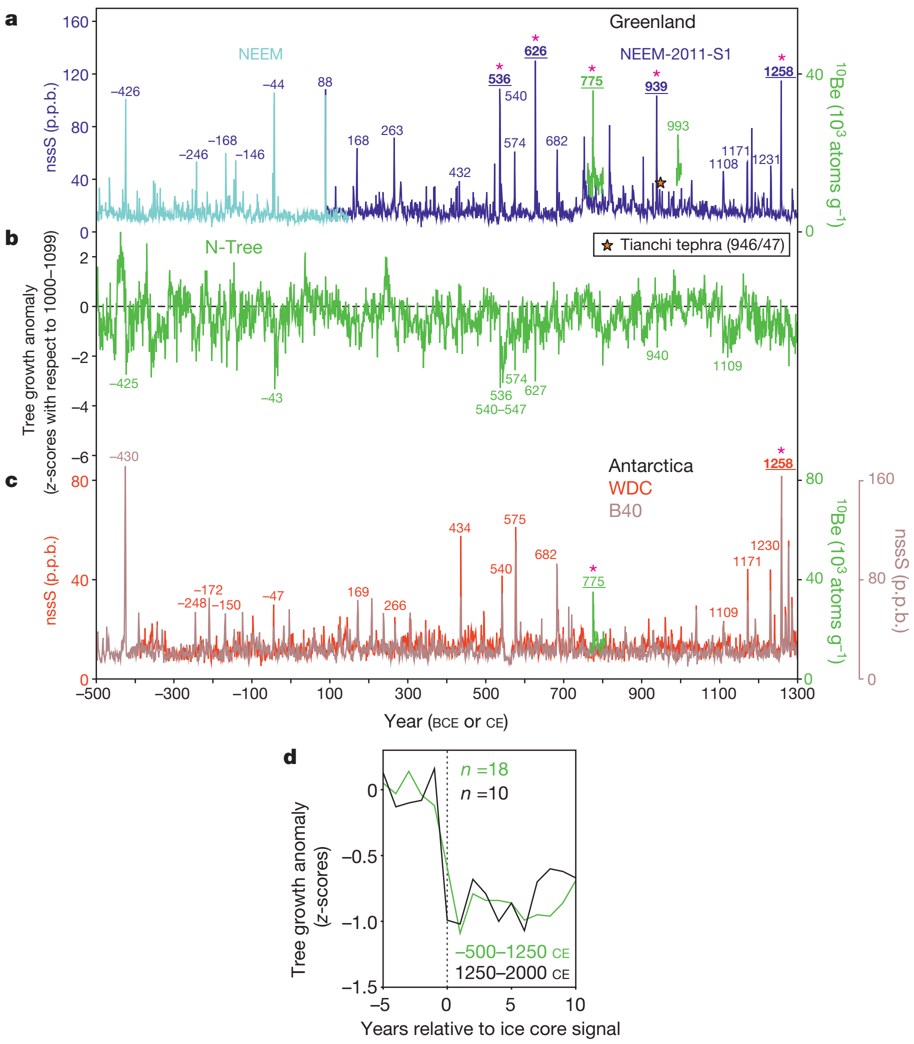


Figure 2 | Re-dated ice-core, non-sea-salt sulfur records from Greenland and Antarctica in relation to growth anomalies in the N-Tree composite. **a**, Ice-core, non-sea-salt sulfur (nssS in parts per billion, p.p.b.) records from Greenland (NEEM, NEEM-2011-S1) on the NS1-2011 timescale between 500 BCE and 1300 CE, with the identified layer of Tianchi tephra[67] highlighted (orange star). Calendar years are given for the start of volcanic sulfate deposition. Events used as fixed age markers to constrain the dating (536 CE, 626 CE, 775 CE, 939 CE and 1258 CE) are indicated (purple stars). Annually resolved ^{10}Be concentration record (green) from NEEM-2011-S1 encompassing the two Δ^{14}C excursion events in trees from 775 CE and 994 CE. **b**, Tree-ring growth anomalies (relative to 1000–1099 CE) for the N-Tree composite[42,43,76–78]. **c**, nssS records from Antarctica (red, WDC; pink, B40) on the WD2014 timescale and annually resolved ^{10}Be concentrations from WDC. **d**, Superposed epoch analysis for 28 large volcanic signals during the past 2,500 years. Tree-ring growth anomalies relative to the timing of reconstructed sulfate deposition in Greenland (NS1-2011) are shown for 1250–2000 CE (black trace) and 500 BCE to 1250 CE (green trace).

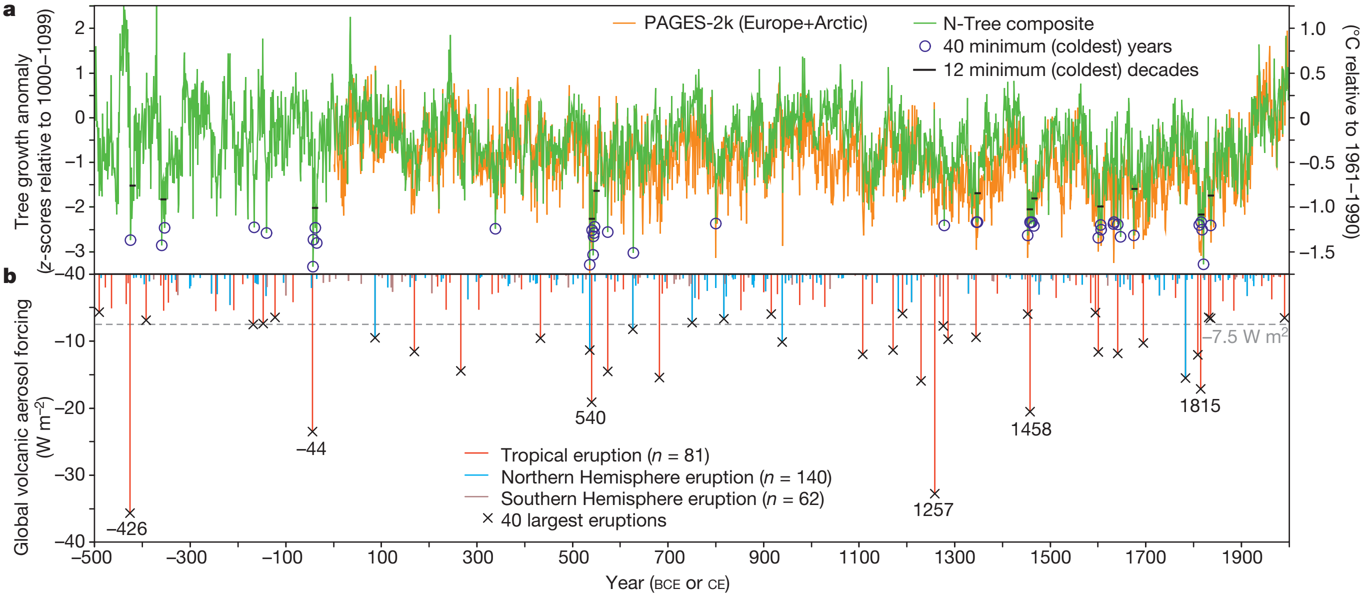


Figure 3 | Global volcanic aerosol forcing and Northern Hemisphere temperature variations for the past 2,500 years. **a**, 2,500-year record of tree-growth anomalies (N-Tree[42,43,76–78]; relative to 1000–1099 CE) and reconstructed summer temperature anomalies for Europe and the Arctic[3] with the 40 coldest single years and the 12 coldest decades based on N-Tree indicated. **b**, Reconstructed global volcanic aerosol forcing from bipolar sulfate composite records from tropical (bipolar), Northern Hemisphere, and Southern Hemisphere eruptions. Total (that is, time-integrated) forcing values are calculated by summing the annual values for the duration of volcanic sulfur deposition. The 40 largest volcanic signals are indicated, and ages are given for events representing atmospheric sulfate loading exceeding that of the Tambora 1815 eruption.

History of volcanic forcing

Employing our revised timescales and new high-resolution, ice-core sulfur measurements, we developed an extended reconstruction of volcanic aerosol deposition since early Roman times for both polar ice sheets, from which we then estimated radiative forcing using established transfer functions[15] (Fig. 3, Methods, Supplementary Data 3–5). This forcing series is characterized by improved dating accuracy, annual resolution, and a larger number of ice-core records in the Antarctic ice-core sulfate composite[17] than in previous reconstructions[6,7]. It spans 2,500 years, allowing investigation of climate–volcano linkages more accurately and earlier than with previous reconstructions. It also provides a perspective on volcanic influences during major historical epochs, such as the growth of Roman imperial power and subsequent decline during the 'Migration Period' (the early part of the first millennium CE) in Europe—times of (1) demographic and economic expansion as well as relative societal stability and (2) political turmoil and population instability, respectively[33]. With improved dating and lower volcanic-sulfate detection limits from the Antarctic array[17], we were able to detect, estimate, and attribute volcanic aerosol loading and forcing from 283 individual eruptive events during this period (Fig. 3).

We attributed about half of these to mid- to high-latitude Northern Hemisphere sources, while 81 were attributed to tropical eruptions (having synchronous sulfate deposition on both polar ice sheets). These tropical volcanic eruptions contributed 64% of total volcanic forcing throughout the period, with five events exceeding the sulfate loading from the 1815 Tambora eruption in Indonesia (Fig. 3, Extended Data Table 4). Events in 426 BCE and 44 BCE rival the great 1257 CE Samalas eruption as the largest sulfate-producing eruptions during this time. These two earlier events have not been widely regarded as large tropical eruptions with potential for strong climate impact[20], owing to the lack of complete and synchronized sulfate records from Greenland and Antarctica. We base the claim that these two eruptions were tropical in origin and caused large radiative perturbations on the observation that ice cores from Greenland and Antarctica record coeval (within their respective age uncertainties) and exceptionally high volcanic sulfate concentrations. Both of these events were followed by strong and persistent growth reduction in tree-ring records[34] (Fig. 2) as is typically observed after large tropical eruptions during the Common Era (Fig. 3).

Post-volcanic summer cooling

Superposed epoch analyses (Methods) performed on the 'N-Tree' composite record centred on the largest volcanic signals between 500 BCE and 1250 CE as well as between 1250 CE and 2000 CE, show a clear post-volcanic cooling signal. For both periods, maximum tree-ring response lagged the date of initial increase of sulfate deposition by one year (Fig. 2), consistent with the response observed if using only

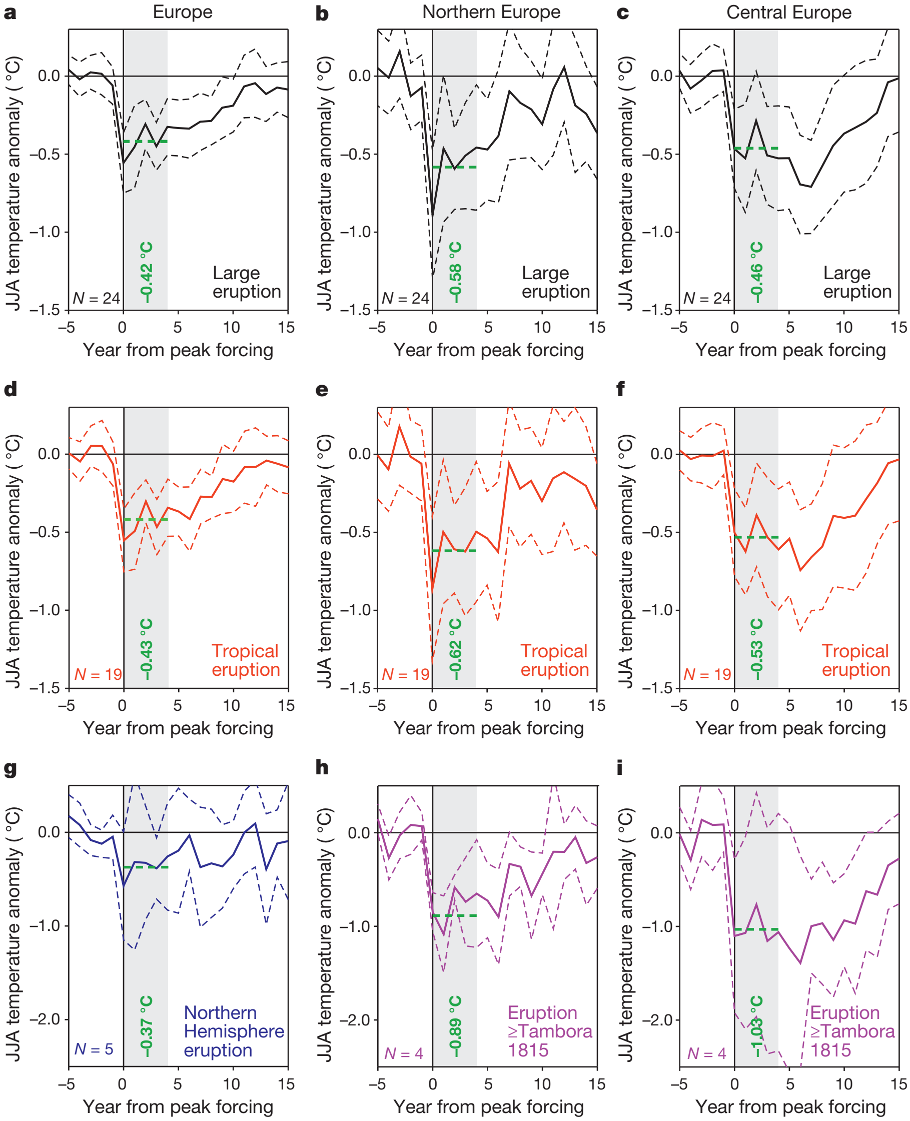


Figure 4 | Post-volcanic cooling. Superposed composites (time segments from selected periods in the Common Era positioned so that the years with peak negative forcing are aligned) of the JJA (June, July and August) temperature response to the 24 largest eruptions (exceeding the Pinatubo 1991 eruption). **a**–**c**, For three regional reconstructions in Europe[3,35,42]. **d**–**f**, For the 19 largest tropical eruptions. **g**, For the five largest Northern Hemisphere eruptions. **h**, **i**, For the eruptions with negative forcing larger than that of the Tambora 1815 eruption for Northern Europe (**h**) and for Central Europe (**i**). Note the different scale for **g**–**i**. JJA temperature anomalies (in °C) for 15 years after reconstructed volcanic peak forcing, relative to the five years before the volcanic eruption, are shown. Dashed lines present twice the standard error of the mean (2 s.e.m.) of the temperature anomalies associated with the multiple eruptions. Five-year average post-volcanic temperatures are shown for each reconstruction (lag 0 to lag +4 years, grey shading).

historically documented eruptions with secure dating for the past 800 years[18]. The sharp and immediate (that is, less than one year lag time) response of tree growth to the ice-core volcanic signal throughout the past 2,500 years further corroborates the accuracy of our new ice-core timescales (Extended Data Fig. 4).

Of the 16 most negative tree-growth anomalies (that is, the coldest summers) between 500 BCE and 1000 CE, 15 followed large volcanic signals—with the four coldest (43 BCE, 536 CE, 543 CE, and 627 CE) occurring shortly after several of the largest events (Extended Data Tables 4 and 5). Similarly, the coldest summers in Europe during the Common Era[3] were associated with large volcanic eruptions (Extended Data Table 5). Reduced tree growth after volcanic eruptions was also prominent in decadal averages of the 'N-Tree' composite. All 16 decades with the most reduced tree growth for our 2,500-year period followed large eruptions (Fig. 3, Extended Data Table 5). In many cases, such as the coldest decade, from 536 CE to 545 CE[3], sustained cooling was associated with the combined effect of several successive volcanic eruptions.

Strong post-volcanic cooling was not restricted to tropical eruptions; it also followed Northern Hemisphere eruptions (Fig. 4), with maximum cooling in the year of volcanic-sulfate deposition. In contrast to the average of the 19 largest CE tropical eruptions, however, the Northern-Hemisphere-only eruptions did not give rise to any noticeable long-term cooling effect (Fig. 4). The persistence of implied post-volcanic cooling following the largest tropical eruptions is strongly expressed in temperature reconstructions based on tree-ring width measurements (for example, those from the Alps), with recovery times of more than ten years. Persistent cooling, with temperature reduction notably below the pre-eruption baseline for six consecutive years, is also observed in temperature reconstructions based on maximum latewood density (for example, those from Northern Scandinavia), which is the preferred proxy with which to quantify volcanic cooling contributions on climate owing to its less marked biological memory effects[35] (Fig. 4). These findings indicate that eruption-induced climate anomalies following large tropical eruptions may last longer than is indicated in many climate simulations (<3–5 years)[9,36,37] and that potential positive feedbacks initiated after large tropical eruptions (for example, sea-ice feedbacks) may not be adequately represented in climate simulations[38,39].

The five-year averaged (lag 0 to lag 4 years) cooling response over three Northern Hemisphere regions (Methods) following the 19 largest Common Era tropical eruptions was −0.6 ± 0.2 °C (two standard errors of the mean, 2 s.e.m.), and that of large Northern Hemisphere eruptions was −0.4 ± 0.4 °C, with the strongest cooling induced in the high latitudes. Overall, cooling was proportional to the magnitude of volcanic forcing, with stratospheric sulfate loading exceeding that of the Tambora eruption inducing the strongest response of −1.1 ± 0.6 °C (Figs 3 and 4).

Global climate anomalies in 536–550 CE

Our new dating allowed us to clarify long-standing debates concerning the origin and consequences of the severe and apparently global climate anomalies observed in the period 536–550 CE, which began with the recognition of the "mystery cloud" of 536 CE[40] observed in the Mediterranean basin. Under previous ice-core dating, it has been argued that this dust veil corresponded to an unknown tropical eruption dated 533–534 CE (±2 years)[41]. Using our revised timescales, we found at least two large volcanic eruptions around this period (Fig. 5).

The first eruptive episode in 535 CE or early 536 CE injected large amounts of sulfate and ash into the atmosphere, apparently in the Northern Hemisphere. Geochemistry of tephra filtered from the NEEM-2011-S1 ice core at a depth corresponding to 536 CE indicated multiple North American volcanoes as likely candidates for a combined volcanic signal (Extended Data Fig. 5, Methods, Supplementary Data 5). Historical observations (Extended Data Table 3) identified atmospheric dimming as early as 24 March 536 CE, and lasting up to 18 months. The summer of 536 CE appeared exceptionally cold in all tree-ring reconstructions in the extra-tropical Northern Hemisphere from North America[34], over Europe[35,42,43] to Asia[44]. Depending on the reconstruction method used, European summer temperatures in 536 CE dropped 1.6–2.5 °C relative to the previous 30-year average[3].

The second eruptive episode in 539 CE or 540 CE, identified in both Greenland and Antarctica ice-core records and hence probably tropical in origin, resulted in up to 10% higher global aerosol loading than

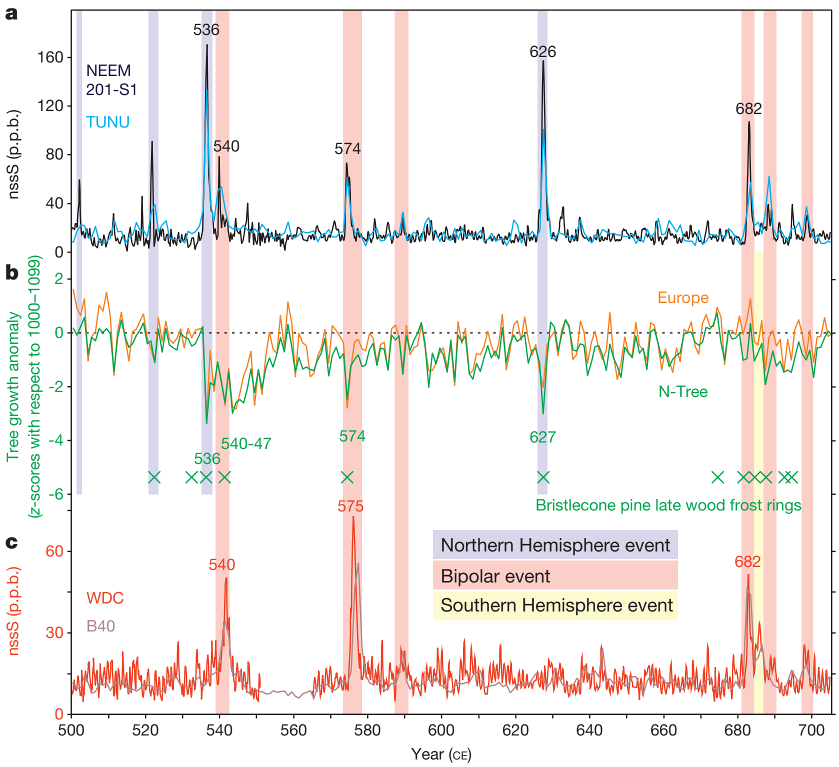


Figure 5 | Volcanism and temperature variability during the migration period (500–705 CE). **a**, Ice-core non-sea-salt sulphur (nssS) records from Greenland (black trace, NEEM-2011-S1; blue trace, TUNU2013). Calendar years for five large eruptions are given for the start of volcanic sulfate deposition. **b**, Summer temperature anomalies (orange trace) for Europe[3], and reconstructed N-Tree growth anomalies (green trace) and occurrence of frost rings in North American bristlecone pine tree-ring records. **c**, nssS records from Antarctica (red trace, WDC; pink trace, B40) on the WD2014 timescale; attribution of the sulfur signals to bipolar, Northern Hemisphere, and Southern Hemisphere events based on the timing of deposition on the two independent timescales is indicated by shading.

the Tambora 1815 eruption reconstructed from our bipolar sulfate records. Summer temperatures consequently dropped again, by 1.4–2.7 °C in Europe in 541 CE[3], and cold temperatures persisted in the Northern Hemisphere until almost 550 CE[3,33,34,42] (Figs 2, 3, 5).

This provides a notable environmental context to widespread famine and the great Justinian Plague of 541–543 CE that was responsible for decimating populations in the Mediterranean and potentially China[45,46]. Although certain climatic conditions (for example, wet summers) have been linked to plague outbreaks in the past[47], a direct causal connection of these two large volcanic episodes and subsequent cooling to crop failures and outbreaks of famines and plagues is difficult to prove[33]. However, the exact delineation of two of the largest volcanic signals—with exceptionally strong and prolonged Northern Hemisphere cooling, written evidence of famines and pandemics, as well as the socio-economic decline observed in Mesoamerica (the "Maya Hiatus"[48]), Europe, and Asia—supports the idea that the latter may be causally associated with volcanically induced climatic extremes.

Detailed study of major volcanic events during the sixth century (Fig. 5) and an assessment of post-volcanic cooling throughout the past 2,500 years using stacked tree-ring records and regional temperature reconstructions (Fig. 4, Extended Data Fig. 4) demonstrated that large eruptions in the tropics and high latitudes were primary drivers of interannual-to-decadal Northern Hemisphere temperature variability. The new ice-core chronologies imply that previous multi-proxy reconstructions of temperature that include ice-core records[2–4] have diminished high- to mid-frequency amplitudes and must be updated to accurately capture the timing and full amplitude of palaeoclimatic variability.

By creating a volcanic forcing index independent of but consistent with tree-ring-indicated cooling, we provide an essential step towards understanding of external forcing on natural climate variability during the past 2,500 years. With the expected detection of additional rapid Δ^{14}C enrichment events from ongoing efforts in annual-resolution ^{14}C tree-ring analyses[49], there will be opportunities to further constrain ice-core dating throughout the Holocene and develop a framework of precisely dated, globally synchronized proxies of past climate variability and external climate forcing.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 November 2014; accepted 6 May 2015.

Published online 8 July 2015.

1. Robock, A. Volcanic eruptions and climate. *Rev. Geophys.* **38,** 191–219 (2000).
2. Hanhijärvi, S., Tingley, M. P. & Korhola, A. Pairwise comparisons to reconstruct mean temperature in the Arctic Atlantic Region over the last 2,000 years. *Clim. Dyn.* **41,** 2039–2060 (2013).
3. PAGES 2k Consortium. Continental-scale temperature variability during the past two millennia. *Nature Geosci.* **6,** 503 (2013).
4. Mann, M. E. *et al.* Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proc. Natl Acad. Sci. USA* **105,** 13252–13257 (2008).
5. Usoskin, I. G. A history of solar activity over millennia. *Living Rev. Sol. Phys* **10,** 1 (2013).
6. Gao, C. C., Robock, A. & Ammann, C. Volcanic forcing of climate over the past 1500 years: an improved ice core-based index for climate models. *J. Geophys. Res.* **113,** http://dx.doi.org/10.1029/2008JD010239 (2008).
7. Crowley, T. J. & Unterman, M. B. Technical details concerning development of a 1200-yr proxy index of global volcanism. *Earth System Sci. Data* **5,** 187–197 (2013).
8. Mann, M. E., Fuentes, J. D. & Rutherford, S. Underestimation of volcanic cooling in tree-ring-based reconstructions of hemispheric temperatures. *Nature Geosci.* **5,** 202–205 (2012).
9. Mann, M. E., Rutherford, S., Schurer, A., Tett, S. F. B. & Fuentes, J. D. Discrepancies between the modeled and proxy-reconstructed response to volcanic forcing over the past millennium: implications and possible mechanisms. *J. Geophys. Res.* **118,** 7617–7627 (2013).
10. Schurer, A. P., Hegerl, G. C., Mann, M. E., Tett, S. F. B. & Phipps, S. J. Separating forced from chaotic climate variability over the past millennium. *J. Clim.* **26,** 6954–6973 (2013).
11. Anchukaitis, K. J. *et al.* Tree rings and volcanic cooling. *Nature Geosci.* **5,** 836–837 (2012).
12. Büntgen, U. *et al.* Extraterrestrial confirmation of tree-ring dating. *Nature Clim. Change* **4,** 404–405 (2014).
13. Esper, J., Büntgen, U., Luterbacher, J. & Krusic, P. J. Testing the hypothesis of post-volcanic missing rings in temperature sensitive dendrochronological data. *Dendrochronologia* **31,** 216–222 (2013).
14. D'Arrigo, R., Wilson, R. & Anchukaitis, K. J. Volcanic cooling signal in tree ring temperature records for the past millennium. *J. Geophys. Res.* **118,** 9000–9010 (2013).
15. Plummer, C. T. *et al.* An independently dated 2000-yr volcanic record from Law Dome, East Antarctica, including a new perspective on the dating of the 1450s CE eruption of Kuwae, Vanuatu. *Clim. Past* **8,** 1929–1940 (2012).
16. Sigl, M. *et al.* A new bipolar ice core record of volcanism from WAIS Divide and NEEM and implications for climate forcing of the last 2000 years. *J. Geophys. Res.* **118,** 1151–1169 (2013).
17. Sigl, M. *et al.* Insights from Antarctica on volcanic forcing during the Common Era. *Nature Clim. Change* **4,** 693–697 (2014).
18. Esper, J. *et al.* European summer temperature response to annually dated volcanic eruptions over the past nine centuries. *Bull. Volcanol.* **75,** 736 (2013).
19. Douglass, D. H. & Knox, R. S. Climate forcing by the volcanic eruption of Mount Pinatubo. *Geophys. Res. Lett.* **32,** L05710 (2005).
20. Baillie, M. G. L. Proposed re-dating of the European ice core chronology by seven years prior to the 7th century AD. *Geophys. Res. Lett.* **35,** L15813 (2008).
21. Baillie, M. G. L. & McAneney, J. Tree ring effects and ice core acidities clarify the volcanic record of the 1st millennium. *Clim. Past* **11,** 105–114 (2015).
22. Miyake, F., Nagaya, K., Masuda, K. & Nakamura, T. A signature of cosmic-ray increase in AD 774–775 from tree rings in Japan. *Nature* **486,** 240–242 (2012).
23. Miyake, F., Masuda, K. & Nakamura, T. Another rapid event in the carbon-14 content of tree rings. *Nature Commun.* **4,** http://dx.doi.org/10.1038/Ncomms2783 (2013).
24. Usoskin, I. G. *et al.* The AD775 cosmic event revisited: the Sun is to blame. *Astron. Astrophys.* **552,** http://dx.doi.org/10.1051/0004-6361/201321080 (2013).
25. Jull, A. J. T. *et al.* Excursions in the ^{14}C record at A. D. 774–775 in tree rings from Russia and America. *Geophys. Res. Lett.* **41,** 3004–3010 (2014).
26. Güttler, D. *et al.* Rapid increase in cosmogenic ^{14}C in AD 775 measured in New Zealand kauri trees indicates short-lived increase in ^{14}C production spanning both hemispheres. *Earth Planet. Sci. Lett.* **411,** 290–297 (2015).
27. Miyake, F. *et al.* Cosmic ray event of AD 774–775 shown in quasi-annual ^{10}Be data from the Antarctic Dome Fuji ice core. *Geophys. Res. Lett.* **42,** 84–89 (2015).
28. Webber, W. R., Higbie, P. R. & McCracken, K. G. Production of the cosmogenic isotopes H-3, Be-7, Be-10, and Cl-36 in the Earth's atmosphere by solar and galactic cosmic rays. *J. Geophys. Res.* **112,** A10106 (2007).
29. Masarik, J. & Beer, J. An updated simulation of particle fluxes and cosmogenic nuclide production in the Earth's atmosphere. *J. Geophys. Res.* **114,** D11103 (2009).
30. Vinther, B. M. *et al.* A synchronized dating of three Greenland ice cores throughout the Holocene. *J. Geophys. Res.* **111,** D13102 (2006).
31. Lavigne, F. *et al.* Source of the great A.D. 1257 mystery eruption unveiled, Samalas volcano, Rinjani Volcanic Complex, Indonesia. *Proc. Natl Acad. Sci. USA* **110,** 16742–16747 (2013).
32. Winstrup, M. *et al.* An automated approach for annual layer counting in ice cores. *Clim. Past* **8,** 1881–1895 (2012).
33. McCormick, M. *et al.* Climate change during and after the Roman Empire: reconstructing the past from scientific and historical evidence. *J. Interdisc. Hist.* **43,** 169–220 (2012).
34. Salzer, M. W. & Hughes, M. K. Bristlecone pine tree rings and volcanic eruptions over the last 5000 yr. *Quat. Res.* **67,** 57–68 (2007).
35. Esper, J., Duthorn, E., Krusic, P. J., Timonen, M. & Buntgen, U. Northern European summer temperature variations over the Common Era from integrated tree-ring density records. *J. Quat. Sci.* **29,** 487–494 (2014).
36. Crowley, T. J. Causes of climate change over the past 1000 years. *Science* **289,** 270–277 (2000).
37. Driscoll, S., Bozzo, A., Gray, L. J., Robock, A. & Stenchikov, G. Coupled Model Intercomparison Project 5 (CMIP5) simulations of climate following volcanic eruptions. *J. Geophys. Res.* **117,** D17105 (2012).
38. Schneider, D. P., Ammann, C. M., Otto-Bliesner, B. L. & Kaufman, D. S. Climate response to large, high-latitude and low-latitude volcanic eruptions in the Community Climate System Model. *J. Geophys. Res.* **114,** D15101 (2009).
39. Zanchettin, D. *et al.* Inter-hemispheric asymmetry in the sea-ice response to volcanic forcing simulated by MPI-ESM (COSMOS-Mill). *Earth Syst. Dyn.* **5,** 223–242 (2014).
40. Stothers, R. B. Mystery cloud of Ad-536. *Nature* **307,** 344–345 (1984).
41. Larsen, L. B. *et al.* New ice core evidence for a volcanic cause of the AD 536 dust veil. *Geophys. Res. Lett.* **35,** L04708 (2008).
42. Büntgen, U. *et al.* 2500 years of European climate variability and human susceptibility. *Science* **331,** 578–582 (2011).
43. Esper, J. *et al.* Orbital forcing of tree-ring data. *Nature Clim. Change* **2,** 862–866 (2012).
44. D'Arrigo, R. *et al.* 1738 years of Mongolian temperature variability inferred from a tree-ring width chronology of Siberian pine. *Geophys. Res. Lett.* **28,** 543–546 (2001).
45. Zhang, Z. B. *et al.* Periodic climate cooling enhanced natural disasters and wars in China during AD 10–1900. *Proc. R. Soc. B* **277,** 3745–3753 (2010).
46. Stothers, R. B. Volcanic dry fogs, climate cooling, and plague pandemics in Europe and the Middle East. *Clim. Change* **42,** 713–723 (1999).

47. Stenseth, N. C. *et al.* Plague dynamics are driven by climate variation. *Proc. Natl Acad. Sci. USA* **103,** 13110–13115 (2006).
48. Dull, R. A. Evidence for forest clearance, agriculture, and human-induced erosion in Precolumbian El Salvador. *Ann. Assoc. Am. Geogr.* **97,** 127–141 (2007).
49. Taylor, R. E. & Southon, J. Reviewing the Mid-First Millennium BC C-14 ''warp'' using C-14/bristlecone pine data. *Nucl. Instrum. Meth. B* **294,** 440–443 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the many people involved in logistics, drill development and drilling, and ice-core processing and analysis in the field and our laboratories. This work was supported by the US National Science Foundation (NSF). We appreciate the support of the WAIS Divide Science Coordination Office (M. Twickler and J. Souney) for collection and distribution of the WAIS Divide ice core; Ice Drilling and Design and Operations (K. Dahnert) for drilling; the National Ice Core Laboratory (B. Bencivengo) for curating the core; Raytheon Polar Services (M. Kippenhan) for logistics support in Antarctica; and the 109th New York Air National Guard for airlift in Antarctica. NEEM is directed and organized by the Center of Ice and Climate at the Niels Bohr Institute and the US NSF, Office of Polar Programs. It is supported by funding agencies and institutions in Belgium (FNRS-CFB and FWO), Canada (NRCan/GSC), China (CAS), Denmark (FIST), France (IPEV, CNRS/INSU, CEA and ANR), Germany (AWI), Iceland (RannIs), Japan (NIPR), Korea (KOPRI), The Netherlands (NWO/ALW), Sweden (VR), Switzerland (SNF), the UK (NERC), and the USA (the US NSF, Office of Polar Programs). We thank B. Nolan, O. Amir, K. D. Pang, M. McCormick, A. Matthews, and B. Rossignol for assistance in surveying and/or interpreting the historical evidence. We thank S. Kuehn for commenting on possible correlations for the tephra. We thank A. Aldahan and G. Possnert for their support in the NGRIP ^{10}Be preparations and measurements at the Department of Earth Sciences and the Tandem laboratory at Uppsala University. We gratefully acknowledge R. Kreidberg for his editorial advice. The following individual grants supported this work: NSF/OPP grants 0839093, 0968391, and 1142166 to J.R.M. for development of the Antarctic ice core records and NSF/OPP grants 0909541, 1023672, and 1204176 to J.R.M. for development of the Arctic ice core records. M.W. was funded by the Villum Foundation. K.C.W. was funded by NSF/OPP grants 0636964 and 0839137. M.C. and T.E.W. were funded by NSF/OPP grants 0839042 and 0636815. F.L. was funded by the Yale Climate and Energy Institute, Initiative for the Science of the Human Past at Harvard, and the Rachel Carson Center for Environment and Society of the Ludwig-Maximilians-Universität (LMU Munich). C.K. was funded by a Marie Curie FP7 Integration Grant within the 7th European Union Framework Programme. M. Salzer was funded by NSF grant ATM 1203749. R.M. was funded by the Swedish Research Council (DNR2013-8421). The division of Climate and Environmental Physics, Physics Institute, University of Bern, acknowledges financial support by the SNF and the Oeschger Centre.

Author Contributions M. Sigl designed the study with input from J.R.M., M.W., G.P., and F.L. The manuscript was written by M. Sigl, M.W., F.L., and J.R.M., with contributions from K.C.W., G.P., U.B., and B.M.V. in interpretation of the measurements. Ice-core chemistry measurements were performed by J.R.M., M. Sigl, O.J.M., N.C., D.R.P. (NEEM, B40, TUNU2013), and by S.S., H.F., R. Mulvaney (NEEM). K.C.W., T.E.W., and M.C. completed ice core ^{10}Be measurements. F.M. and R. Muscheler were responsible for the NGRIP ice core ^{10}Be measurements. M. Sigl, M.W., B.M.V., and J.R.M. analysed ice-core data and developed age models. F.L. and C.K. analysed historical documentary data. G.P. and J.R.P. performed ice-core tephra analysis and data interpretation. U.B. and M. Salzer contributed tree-ring data. D.D.-J., B.M.V., J.P.S., S.K., and O.J.M. were involved in drilling of the NEEM ice core. TUNU2013 was drilled by M. Sigl, N.C. and O.J.M., and the B40 ice core was drilled by S.K. and made available for chemistry measurements. D.D.-J. and J.P.S. were responsible for NEEM project management, sample distribution, logistics support, and management. All authors contributed towards improving the final manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.R.M. (joe.mcconnell@dri.edu).

METHODS

Ice cores. This study included new and previously described ice-core records from five drilling sites (Extended Data Fig. 1, Supplementary Data 1). The upper 577 m of the 3,405-m WAIS Divide (WDC) core from central West Antarctica and a 410-m intermediate-length core (NEEM-2011-S1) drilled in 2011 close to the 2,540-m North Greenland Eemian Ice Drilling (NEEM)[50] ice core have previously been used to reconstruct sulfate deposition in both polar ice sheets[16]. These coring sites are characterized by relatively high snowfall ($\sim$200 kg m^{-2} yr^{-1}) and have comparable elevation, latitude, and deposition regimes. WDC and NEEM-2011-S1 provided high-resolution records that allowed annual-layer dating based on seasonally varying impurity content[16]. New ice-core analyses included the upper 514 m of the main NEEM core used to extend the record of NEEM-2011-S1 to cover the past 2,500 years, as well as B40 drilled in 2012 in Dronning Maud Land in East Antarctica and TUNU2013 drilled in 2013 in Northeast Greenland—both characterized by lower snowfall rates ($\sim$70–100 kg m^{-2} yr^{-1}). Volcanic sulfate concentration from B40 had been reported previously for the past 2,000 years[17], but we extended measurements to 200 m depth to cover the past 2,500 years.

High-resolution, ice-core aerosol analyses. Ice-core analyses were performed at the Desert Research Institute (DRI) using 55–100-cm-long, longitudinal ice-core sections (33 mm × 33 mm wide). The analytical system for continuous analysis included two Element2 (Thermo Scientific) high-resolution inductively coupled plasma mass spectrometers (HR-ICP-MS) operating in parallel for measurement of a broad range of $\sim$35 elements; an SP2 (Droplet Measurement Technologies) instrument for black carbon measurements; and a host of fluorimeters and spectrophotometers for ammonium (NH_4^+), nitrate (NO_3^-), hydrogen peroxide (H_2O_2), and other chemical species. All measurements were exactly co-registered in depth, with depth resolution typically less than 10–15 mm[51–53]. We corrected total sulfur (S) concentrations for the sea-salt S contribution using sea-salt Na concentrations[16]. Measurements included TUNU2013 and NEEM (400–515 m) in Greenland, and B40 in Antarctica (Extended Data Fig. 1). Gaps (that is, ice not allocated to DRI) in the high-resolution sulfur data of the NEEM core were filled with $\sim$4-cm-resolution discrete sulfate measurements using fast ion-chromatography techniques[54] performed in the field between 428 m and 506 m depth.

Independent analyses of the upper part of the NEEM main core were performed in the field using a continuous flow analysis (CFA) system[55] recently modified to include a new melter head design[56]. Ca^{2+}, NH_4^+, and H_2O_2 were analysed by fluorescence spectroscopy; Na^+ and NO_3^- by absorption spectroscopy; conductivity of the meltwater by a micro flow cell (Amber Science); and a particle detector (Abakus, Klotz) was used for measuring insoluble dust particle concentrations and size distribution[57]. Effective depth resolution was typically better than 20 mm. Measurements were exactly synchronized in depth using a multicomponent standard solution; the accuracy of the depth assignment for all measurements was typically better than 5 mm.

High-resolution measurements of ^{10}Be in ice cores using AMS. Accelerator mass spectrometry (AMS) was used to analyse samples from the NEEM-2011-S1, WDC, NGRIP, and TUNU2013 ice cores encompassing the time period of the Δ^{14}C anomalies from tree-ring records[12,22–25] were used for ^{10}Be analysis (Supplementary Data 1). NEEM-2011-S1 and WDC were sampled in exact annual resolution, using the maxima (minima in WDC) of the annual cycles of Na concentrations to define the beginning of the calendar year[16]. NGRIP was sampled at a constant resolution of 18.3 cm, providing an age resolution of about one year. Similarly, TUNU2013 was sampled in quasi-annual resolution according to the average annual-layer thickness expected at this depth based on prior volcanic synchronization to NEEM-2011-S1. The relative age uncertainty for TUNU2013 with respect to the dependent NEEM-2011-S1 chronology at this depth is assumed to be ±1 year at most, given a distinctive match for selected volcanic trace elements in both ice-core records (752–764 CE, NS1-2011 timescale). Sample masses ranged between 100 g and 450 g, resulting in median overall quantification uncertainties of less than 4%–7%. The ^{10}Be/^{9}Be ratios of samples and blanks were measured relative to well documented ^{10}Be standards[13] by AMS at Purdue's PRIME laboratory (WDC, NEEM-2011-S1, Tunu2013) and Uppsala University (NGRIP)[58,59]. Results were corrected for an average blank ^{10}Be/^{9}Be ratio, corresponding to corrections of 2%–10% of the measured ^{10}Be/^{9}Be ratios.

Annual-layer dating using the StratiCounter algorithm. For annual-layer interpretation, we used DRI's broad-spectrum aerosol concentration data from WDC (188–577 m), NEEM-2011-S1 (183–411 m), and NEEM (410–515 m), as well as NEEM aerosol concentration data (183–514 m) from the field-based CFA system. The original timescale for NEEM-2011-S1 was based on volcanic synchronization to the NGRIP sulfate record on the GICC05 timescale and annual-layer interpretation between the volcanic age markers, whereas WDC was previously dated by annual-layer counting[16].

Parameters with strong intra-annual variability included tracers of sea salt (for example, Na, Cl, Sr), dust (for example, Ce, Mg, insoluble particle concentration), and marine biogenic emissions such as non-sea-salt sulfur (nssS). Tracers of biomass-burning emissions, such as BC, NH_4^+, and NO_3^-, also showed strong seasonal variations in deposition during pre-industrial times[16,60,61]. The data sets used for annual-layer interpretation are provided in Extended Data Table 1. For NEEM-2011-S1, the final database used for annual-layer dating included 13 parameters and the ratio of nssS/Na. For WDC, the final database included five parameters and the ratio of nssS/Na. For NEEM (410–515 m depth), the final database included eight parameters (Na^+, Ca^{2+}, NH_4^+, H_2O_2, NO_3^-, conductivity, insoluble particle concentrations, and electrical conductivity[62]) from the field-based measurements and eleven parameters (Na, Cl, Mg, Mn, Sr, nssS, nssS/Na, nssCa, black carbon, NO_3^-, NH_4^+) from the DRI system.

We focused here on the time period before the large volcanic eruption of Samalas in 1257 CE[31], clearly detectable as an acidic peak in both ice-core records, and consequently started annual-layer counting of NEEM-2011-S1, NEEM, and WDC at the depth of the corresponding sulfur signal. For the time period 1257 CE to present, ice-core chronologies were constrained by numerous historic eruptions and large sulfate peaks, showing a strong association to Northern Hemisphere cooling events as indicated by tree-ring records[16].

We applied the StratiCounter layer-detection algorithm[32] to the multi-parameter aerosol concentration records ($n = 14$ for NEEM-2011-S1; $n = 6$ for WDC; $n = 8$ for NEEM $<$ 410 m; $n = 19$ for NEEM $>$ 410 m) to objectively determine the most likely number of annual layers in the ice cores along with corresponding uncertainties. The StratiCounter algorithm is based on statistical inference in Hidden Markov Models (HMMs), and it determines the maximum-likelihood solution based on the annual signal in all aerosol records in parallel. Some of these displayed a high degree of similarity, so we weighted these records correspondingly lower. The algorithm was run step-wise down the core, each batch covering approximately 50 years, with a slight overlap. All parameters for the statistical description of a mean layer and its inter-annual variability in the various aerosol records were determined independently for each batch as the maximum-likelihood solution. The algorithm simultaneously computes confidence intervals for the number of layers within given sections, allowing us to provide uncertainty bounds on the number of layers between selected age-marker horizons (Extended Data Table 2).

Annual-layer detection in the NEEM main core below 410 m was made more difficult by frequent occurrence of small gaps in the two independent high-resolution aerosol data sets. Depending on the parameter, data gaps from the CFA field measurements accounted for up to 20% of the depth range between 410 m and 515 m, but the combined aerosol records from both analyses provided an almost complete aerosol record with 96% data coverage. As this was the first time that the StratiCounter algorithm was used simultaneously on data records from two different melt systems, with different characteristics and lack of exact co-registration, we also manually determined annual layers below 410 m using the following approaches: one investigator used Na and nssCa concentrations and the ratio of nssS/Na (from DRI analysis) as well as Na^+ and insoluble particle concentrations (from CFA analysis) as primary dating parameters. Black carbon, NH_4^+, nssS, and conductivity were used as secondary dating parameters where annual-layer interpretation was ambiguous. A second investigator used DRI's Na, Ca, BC, NH_4^+ and CFA Na^+, Ca^{2+}, and NH_4^+ measurements as parameters. The annual-layer interpretation of the NEEM core between 410 m and 514 m from investigator 1 was within the interpretation uncertainties of the StratiCounter output, from which it differed by less than a single year over the majority of this section, and it differed from independently counted timescales (for example, GICC05)[62] by on average less than three years (Extended Data Fig. 2). This set of layer counts was used for the resulting timescale.

New ice-core chronologies for NS1-2011 and WD2014. We defined the depth of NEEM-2011-S1 containing the maximum ^{10}Be concentration as the year 775 CE. Relative to this constraint, the maximum-likelihood ages for three large volcanic sulfate peaks were within one year of documented historical reports from early written sources of prominent and sustained atmospheric dimming observed in Europe and/or the Near East (Extended Data Table 3, Supplementary Data 2). Automated-layer identification for NEEM-2011-S1 was therefore constrained by tying the respective ice-core volcanic signals to the corresponding absolute historically dated ages of 536 CE, 626 CE, and 939 CE (Extended Data Table 2)—thereby creating a new ice-core timescale (NS1-2011). The volcanic sulfur signal corresponding to the eruption of Samalas believed to have occurred in late 1257[31] was constrained to 1258 CE to account for several months' delay in sulfate deposition in the high latitudes. Before 86 CE (the bottom depth of NEEM-2011-S1), the NS1-2011 timescale was extended using the manually derived annual-layer interpretation of the combined NEEM aerosol data sets back to 500 BCE (Fig. 2).

In NS1-2011 we did not attribute acid layers to the historical eruptions Vesuvius 79 and Hekla 1104, due to a lack of corroborative tephra at these depths in this and a previous study[63]. Possible Vesuvian tephra was reported from the Greenland Ice Sheet Project (GRIP) ice core at 429.3 m depth[64], but in view of the new annual-layer dating results (Extended Data Fig. 3), we concluded that this layer dates to 87/88 CE. Furthermore, volcanic sulfate deposition values for the corresponding event show a strong spatial gradient over Greenland with highest values in north-west Greenland[16] and lowest in central and south Greenland[65], favouring the attribution of a volcanic source from the high latitudes. Documentary sources (Supplementary Data 2) also suggest that the main vector of ash transport following the Vesuvius 79 CE eruption was towards the eastern Mediterranean[66].

For WDC, we do not have other sufficiently well determined age constraints besides the rapid ^{10}Be increase in 775 CE and the sulfur signal of the Samalas 1257 eruption. Therefore, no additional constraints were used when creating the new ice-core timescale ("WD2014") from the StratiCounter annual-layer interpretation back to 396 BCE.

Depth-age information for six distinctive marker horizons in Greenland is given, and five of these horizons were used to constrain NS1-2011 (Extended Data Table 3). Similarly, depth information, the number of annual layers, and 95% confidence intervals between distinctive volcanic marker horizons are given for NEEM, NEEM-2011-S1, and WDC, supporting attribution of these ice-core signals to eruptions in the low latitudes with bipolar sulfate deposition.

Evaluation of NS1-2011 using independent age information. We evaluated timescale accuracy using additional distinctive age markers not used during chronology development:

(1) Tephra from the eruption of Changbaishan/Tianchi (China)[67] was detected in NEEM-2011-S1 in 946–947 CE, in agreement with widespread documentary evidence of an eruption in that region in winter 946/47 CE[68] also supported by a high-precision ^{14}C wiggle-match age of 946 ± 3 CE obtained from a tree killed during this eruption[68].

(2) The rapid increase of ^{10}Be from the 994 CE event occurred in NEEM-2011-S1 in 993 CE, consistent with Δ^{14}C from Japanese tree rings showing that the rapid increase in radionuclide production took place between the Northern Hemisphere growing seasons of 993 CE and 994 CE[23].

(3) To assess the accuracy of the NS1-2011 timescale before the earliest age marker at 536 CE, we compiled an independent time series of validation points, featuring years with well dated historical reports of atmospheric phenomena associated with high-altitude volcanic dust and/or aerosols (Supplementary Data 2) as known from modern observations to occur after major eruptions (for example, the Krakatau eruption of 1883). These phenomena include diminished sunlight, discoloration of the solar disk, solar coronae (that is, Bishop's Rings), and deeply red twilights (that is, volcanic sunsets)[69,70]. Thirty-two events met our criteria as validation points for the pre-536 CE NS1-2011 timescale. For the earliest in 255 BCE, it was reported in Babylon that "the disk of the sun looked like that of the moon"[71]. For the latest in 501 CE, it was reported in North China that "the Sun was red and without brilliance"[72]. We found that NEEM volcanic event years (including both NEEM and NEEM-2011-S1 data) occurred closely in time (that is, within a conservative ±3-year margin) to 24 (75.0%) of our validation points (Extended Data Fig. 2). To assess whether this association arose solely by chance, we conducted a Monte Carlo equal means test with 1,000,000 iterations (Supplementary Data 2) and found that the number of volcanic event years within three years of our validation points was significantly greater than expected randomly ($P < 0.001$). A significant association was also observed ($P < 0.001$) when using less conservative error margins (±1 and ±2 years) and when excluding any historical observations with less certainty of a volcanic origin (Supplementary Data 2). When placing volcanic event years on the original GICC05 timescale, we did not observe any statistically significant association with our independent validation points.

Potential causes of a previous ice-core dating bias. Interpretation of annual layers in ice cores is subject to accumulating age uncertainty due to ambiguities in the underlying ice-core profiles[30,73]. Bias in existing chronologies may arise from several factors, including: (1) low effective resolution of some ice-core measurements (NGRIP, GRIP); (2) use of only single (or few) parameters for annual-layer interpretation (GRIP, Dye-3 ice cores); (3) intra-annual variations in various ice-core parameters falsely interpreted as layer boundaries (for example, caused by summer melt in Dye-3)[74]; (4) use of tephra believed to originate from the 79 CE Vesuvian eruption[64] as a fixed reference horizon to constrain the Greenland ice-core dating[30]; (5) use of manual-layer interpretation techniques that may favour interpretations consistent with a priori knowledge or existing chronologies (WDC)[16,21].

Volcanic synchronization of B40, TUNU2013, and NGRIP. Two high-resolution sulfur ice-core records (TUNU2013, Greenland and B40, Antarctica) were synchronized to NEEM-2011-S1 and WDC, respectively, using volcanic stratigraphic age markers[17] with relative age uncertainty between the tie-points estimated to not exceed ±2 years. The NGRIP sulfate record measured at 5 cm depth resolution[15] similarly was synchronized to NS1-2011 using 124 volcanic tie-points between 226 and 1999 CE. During the time period with no sulfur record yet available for WDC (before 396 BCE), a tentative chronology for B40 was derived by linearly extrapolating mean annual-layer thickness for B40 as derived from the synchronization to WDC between the earliest volcanic match points.

2,500 year global volcanic forcing ice-core index. We constructed an index of global volcanic aerosol forcing by (1) re-dating and extending to 500 BCE an existing reconstruction of sulfate flux from an Antarctic ice-core array[17] by applying an area weighting of 80/20 between East Antarctica and West Antarctica to B40 and WDC volcanic sulfate flux values, respectively; (2) compositing NGRIP and the NEEM-2011-S1/NEEM sulfate flux records to a similar Greenland sulfate deposition composite back to 500 BCE; (3) using established scaling functions[6,75] to estimate hemispheric sulfate aerosol loading from both polar ice-core composites; and (4) scaling global aerosol loading to the total (that is, time-integrated) radiative volcanic aerosol forcing following the Tambora 1815 eruption[7]. Since the NS1-2011 and WD2014 timescales are independent of each other, the timing of bipolar events had to be adjusted to follow a single timescale to derive a unified global volcanic forcing series. We chose NS1-2011 as the reference chronology for most of the volcanic time series because this age model was constrained and validated by more stratigraphic age markers than WD2014. WD2014 was used as the reference chronology only between 150 CE and 450 CE, because of better data quality during that time period. TUNU2013 was not included in the Greenland ice-core composite because annual-layer thickness variability at this site is influenced strongly by glaciological processes, leading to relatively large uncertainties in atmospheric sulfur-deposition determinations.

Northern Hemisphere tree-ring composite. Tree-ring records from certain locations reflect summer cooling (as is widely observed after volcanic eruptions) with no age uncertainty in annual ring-width dating, thus allowing independent validation of ice-core timescales and the derived volcanic forcing indices. However, no tree-ring-based temperature reconstructions of large spatial scales span the full 2,500 years represented by our new ice-core chronologies. To thus evaluate our new ice-core chronologies and assess the consistency of response throughout the past 2,500 years, we compiled a composite (entitled 'N-Tree') of multi-centennial tree growth records at locations where temperature is the limiting growth factor. We selected available Northern Hemisphere tree-ring records that provided a continuous record of >1,500 years and showed a significant positive relationship with JJA temperatures during the instrumental period (1901–2000 CE) with $P < 0.005$ (adjusted for a reduced sample size owing to autocorrelation of the data sets). In total, five tree-ring chronologies (three based on ring-width measurements, two based on measurements of maximum latewood density) met these criteria[42,43,76–78] of which three are located in the high latitudes of Eurasia (Extended Data Fig. 1).

As various climatic and non-climatic parameters may influence sensitivity of tree growth to temperatures during the twentieth century[79–81], we used the time period 1000-1099 CE as a common baseline for standardizing tree growth anomalies among the five chronologies and built a tree growth composite record called N-Tree (z-scores) by averaging the individual records. Correlations between N-Tree ($N = 5$) and the average of three regional reconstructions for the Arctic, Europe, and Asia ($N > 275$)[3] between 1800 CE and 2000 CE are very high ($r = 0.86$, $N = 201$, $P < 0.0001$), suggesting that much of the large-scale variation in temperature is explained by these selected tree-ring records. Three records in N-Tree cover the period from 138 BCE to the present, thus allowing at least a qualitative assessment of the coherence of growth reduction following large volcanic eruptions before the Common Era (Fig. 2, Extended Data Fig. 4).

Temperature reconstructions. To quantify the Common Era climate impact and investigate regional differences, we used tree-ring-based JJA temperature reconstructions covering the past 2,000 years with a demonstrated strong relationship ($r \geq 0.45$; $P < 0.0001$; Extended Data Fig. 1) to instrumental JJA temperature data[82] between 1901 and 2000. For regions where this criterion was met by several reconstructions (for example, Scandinavia), we limited the analysis to the most recently updated reconstruction[35]. Three regional reconstructions from Central Europe[42], Northern Europe[35], and Northern Siberia (Yamal, not shown)[76] as well as a continental-scale reconstruction for Europe[3] met this criterion and were used to quantify the average response of summer temperature to volcanic forcing during the Common Era (Figs 3 and 4).

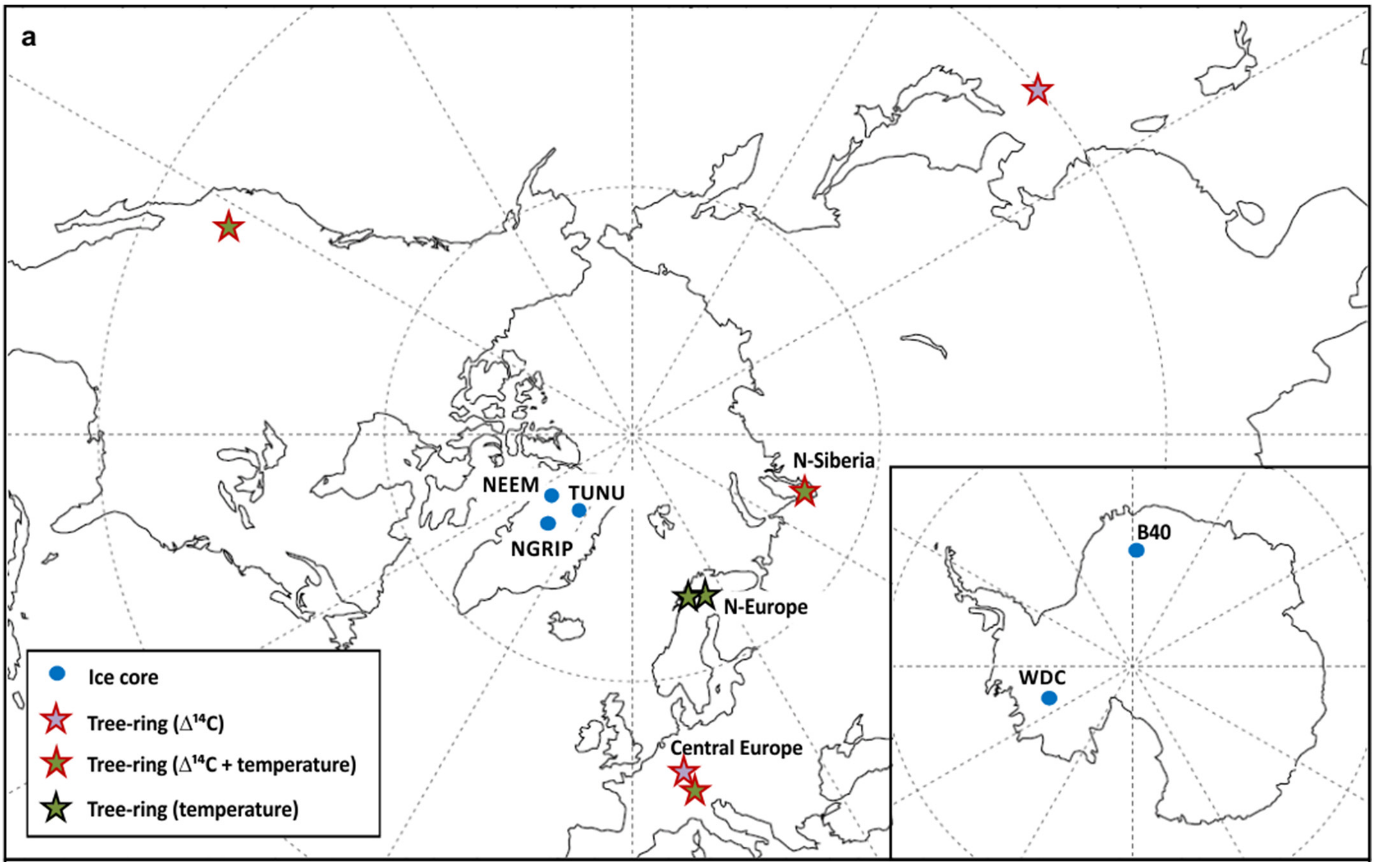
Superposed epoch analyses. To assess tree-ring growth reduction and summer cooling following large eruptions, we used superposed epoch analyses[83,84]. We selected all volcanic eruptions (28 events in total, 24 CE events) with time-integrated volcanic forcing greater than $-7.5\,\mathrm{W\,m^{-2}}$ (that is, eruptions larger than Pinatubo 1991) and aligned the individual segments of N-Tree and regional

JJA temperature reconstructions relative to ice-core-indicated peak forcing. The composite response was calculated for the average of the individual series (lag 0 to lag 10 or 15 years) relative to the average values five years before individual volcanic events (lag −5 to lag −1 year). 95% confidence intervals represent 2 s.e.m. of the tree-growth (Extended Data Fig. 4) and temperature anomalies (Fig. 4) associated with the multiple eruptions.

Cryptotephra analyses of the 536 CE sample from NEEM-2011-S1. We analysed samples from NEEM-2011-S1 for tephra between 326.73 m and 328.06 m depth, corresponding to 531–539 CE (NS1-2011 timescale). Samples (200 g to 500 g) were filtered, and elemental composition of recovered volcanic glass shards determined by electron microprobe analysis at Queen's University Belfast using established protocols[63,67,85] and secondary glass standards[86,87]. Between 326.73 m and 327.25 m, large volume samples were cut at 8 cm depth resolution (≤0.5 years) and with an average cross-section of 26 cm^2. Between 327.25 m and 328.06 m, the average cross-section was 7 cm^2 and depth resolution 20 cm (~1 yr resolution). Tephra particles ($n \geq 17$) were isolated from a sample of ice (327.17–327.25 m depth, 251 g) corresponding to the sulfate spike at 536 CE. The glass shards were heterogeneous in size (20–80 μm), morphology (platey, blocky, vesicular, microlitic), and geochemistry (andesitic, trachytic, rhyolitic). Individual shards had geochemical compositions that share affinities with volcanic systems in the Aleutian arc (Alaska)[88], Northern Cordilleran volcanic province (British Columbia)[89], and Mono-Inyo Craters area (California)[90,91]—indicating at least three synchronous eruptive events, all situated in western North America between 38 °N and 58 °N (Extended Data Fig. 5; Supplementary Data 5).

Data and code availability. Ice-core data (chemistry, including sulphur and ^{10}Be), the resulting timescales, and the volcanic forcing reconstruction are provided as Supplementary Data 1, and 3–5). Historical documentary data are provided as Supplementary Data 2. The code for the StratiCounter program is accessible at the github repository (http://www.github.com/maiwinstrup/StratiCounter). NGRIP SO_4 data can be obtained at http://www.iceandclimate.nbi.ku.dk/data/2012-12-03_NGRIP_SO4_5cm_Plummet_et_al_CP_2012.txt. Tree-ring records and temperature reconstructions are from the Supplementary Database S1 and S2 of the Pages-2k Consortium (ref. 3; http://www.nature.com/ngeo/journal/v6/n5/full/ngeo1797.html#supplementary-information).

50. Dahl-Jensen, D. *et al.* Eemian interglacial reconstructed from a Greenland folded ice core. *Nature* **493,** 489–494 (2013).
51. McConnell, J. R. Continuous ice-core chemical analyses using inductively coupled plasma mass spectrometry. *Environ. Sci. Technol.* **36,** 7–11 (2002).
52. McConnell, J. R. & Edwards, R. Coal burning leaves toxic heavy metal legacy in the Arctic. *Proc. Natl Acad. Sci. USA* **105,** 12140–12144 (2008).
53. Pasteris, D. R. *et al.* Seasonally resolved ice core records from West Antarctica indicate a sea ice source of sea-salt aerosol and a biomass burning source of ammonium. *J. Geophys. Res.* **119,** 9168–9182 (2014).
54. Abram, N. J., Mulvaney, R. & Arrowsmith, C. Environmental signals in a highly resolved ice core from James Ross Island, Antarctica. *J. Geophys. Res.* **116,** D20116 (2011).
55. Kaufmann, P. R. *et al.* An improved continuous flow analysis system for high-resolution field measurements on ice cores. *Environ. Sci. Technol.* **42,** 8044–8050 (2008).
56. Bigler, M. *et al.* Optimization of High-Resolution Continuous Flow Analysis for Transient Climate Signals in Ice Cores. *Environ. Sci. Technol.* **45,** 4483–4489 (2011).
57. Ruth, U., Wagenbach, D., Steffensen, J. P. & Bigler, M. Continuous record of microparticle concentration and size distribution in the central Greenland NGRIP ice core during the last glacial period. *J. Geophys. Res.* **108** (2003).
58. Woodruff, T. E., Welten, K. C., Caffee, M. W. & Nishiizumi, K. Interlaboratory comparison of Be-10 concentrations in two ice cores from Central West Antarctica. *Nucl. Instrum. Meth. B* **294,** 77–80 (2013).
59. Berggren, A. M. *et al.* Variability of Be-10 and delta O-18 in snow pits from Greenland and a surface traverse from Antarctica. *Nucl. Instrum. Meth. B* **294,** 568–572 (2013).
60. Bisiaux, M. M. *et al.* Changes in black carbon deposition to Antarctica from two high-resolution ice core records, 1850-2000 AD. *Atmos. Chem. Phys.* **12,** 4107–4115 (2012).
61. Pasteris, D., McConnell, J. R., Edwards, R., Isaksson, E. & Albert, M. R. Acidity decline in Antarctic ice cores during the Little Ice Age linked to changes in atmospheric nitrate and sea salt concentrations. *J. Geophys. Res.* **119,** 5640–5652 (2014).
62. Rasmussen, S. O. *et al.* A first chronology for the North Greenland Eemian Ice Drilling (NEEM) ice core. *Clim. Past* **9,** 2713–2730 (2013).
63. Coulter, S. E. *et al.* Holocene tephras highlight complexity of volcanic signals in Greenland ice cores. *J. Geophys. Res.* **117,** D21303 (2012).
64. Barbante, C. *et al.* Greenland ice core evidence of the 79 AD Vesuvius eruption. *Clim. Past* **9,** 1221–1232 (2013).
65. Clausen, H. B. *et al.* A comparison of the volcanic records over the past 4000 years from the Greenland Ice Core Project and Dye 3 Greenland Ice Cores. *J. Geophys. Res.* **102,** 26707–26723 (1997).
66. Rolandi, G., Paone, A., Di Lascio, M. & Stefani, G. The 79 AD eruption of Somma: the relationship between the date of the eruption and the southeast tephra dispersion. *J. Volcanol. Geotherm. Res.* **169,** 87–98 (2008).
67. Sun, C. Q. *et al.* Ash from Changbaishan millennium eruption recorded in Greenland ice: implications for determining the eruption's timing and impact. *Geophys. Res. Lett.* **41,** 694–701 (2014).
68. Xu, J. D. *et al.* Climatic impact of the millennium eruption of Changbaishan volcano in China: new insights from high-precision radiocarbon wiggle-match dating. *Geophys. Res. Lett.* **40,** 54–59 (2013).
69. Deirmendjian, D. On volcanic and other particulate turbidity anomalies. *Adv. Geophys.* **16,** 267–296 (1973).
70. Vollmer, M. Effects of absorbing particles on coronas and glories. *Appl. Opt.* **44,** 5658–5666 (2005).
71. Sachs, A. J. & Hunger, H. *Astronomical Diaries and Related Texts from Babylonia* Vol.3 *Diaries from 164 B.C. to 61 B.C.* (Verlag der Österreichischen Akademie der Wissenschaften, 1996).
72. Wittmann, A. D. & Xu, Z. T. A catalog of sunspot observations from 165 BC to AD 1684. *Astron. Astrophys.* (Suppl.) **70,** 83–94 (1987).
73. Rasmussen, S. O. *et al.* A new Greenland ice core chronology for the last glacial termination. *J. Geophys. Res.* **111,** D06102 (2006).
74. Herron, M. M., Herron, S. L. & Langway, C. C. Climatic signal of ice melt features in southern Greenland. *Nature* **293,** 389–391 (1981).
75. Gao, C. H., Oman, L., Robock, A. & Stenchikov, G. L. Atmospheric volcanic loading derived from bipolar ice cores: accounting for the spatial distribution of volcanic deposition. *J. Geophys. Res.* **112,** D09109 (2007).
76. Briffa, K. R. *et al.* Reassessing the evidence for tree-growth and inferred temperature change during the Common Era in Yamalia, northwest Siberia. *Quat. Sci. Rev.* **72,** 83–107 (2013).
77. Grudd, H. Tornetrask tree-ring width and density AD 500-2004: a test of climatic sensitivity and a new 1500-year reconstruction of north Fennoscandian summers. *Clim. Dyn.* **31,** 843–857 (2008).
78. Salzer, M. W., Bunn, A. G., Graham, N. E. & Hughes, M. K. Five millennia of paleotemperature from tree-rings in the Great Basin, USA. *Clim. Dyn.* **42,** 1517–1526 (2014).
79. McMahon, S. M., Parker, G. G. & Miller, D. R. Evidence for a recent increase in forest growth. *Proc. Natl Acad. Sci. USA* **107,** 3611–3615 (2010).
80. Salzer, M. W., Hughes, M. K., Bunn, A. G. & Kipfmueller, K. F. Recent unprecedented tree-ring growth in bristlecone pine at the highest elevations and possible causes. *Proc. Natl Acad. Sci. USA* **106,** 20348–20353 (2009).
81. Briffa, K. R. *et al.* Reduced sensitivity of recent tree-growth to temperature at high northern latitudes. *Nature* **391,** 678–682 (1998).
82. Rohde, R. *et al.* A new estimate of the average land surface temperature spanning 1753 to 2011. *Geoinform. Geostat. Overview* **1,** http://dx.doi.org/10.4172/2327-4581.1000101 (2013).
83. Mass, C. F. & Portman, D. A. Major volcanic eruptions and climate: a critical evaluation. *J. Clim.* **2,** 566–593 (1989).
84. Fritts, H. C., Lofgren, G. R. & Gordon, G. A. Variations in climate since 1602 as reconstructed from tree rings. *Quat. Res.* **12,** 18–46 (1979).
85. Jensen, B. J. L. *et al.* Transatlantic distribution of the Alaskan White River Ash. *Geology* **42,** 875–878 (2014).
86. Oskarsson, N., Sigvaldason, G. E. & Steinthorsson, S. A dynamic-model of rift-zone petrogenesis and the regional petrology of Iceland. *J. Petrol.* **23,** 28–74 (1982).
87. Kuehn, S. C., Froese, D. G., Shane, P. A. R. & Participants, I. I. The INTAV intercomparison of electron-beam microanalysis of glass by tephrochronology laboratories: results and recommendations. *Quat. Int.* **246,** 19–47 (2011).
88. Kaufman, D. S. *et al.* Late Quaternary tephrostratigraphy, Ahklun mountains, SW Alaska. *J. Quat. Sci.* **27,** 344–359 (2012).
89. Lakeman, T. R. *et al.* Holocene tephras in lake cores from northern British Columbia, Canada. *Can. J. Earth Sci.* **45,** 935–947 (2008).
90. Bursik, M., Sieh, K. & Meltzner, A. Deposits of the most recent eruption in the Southern Mono Craters, California: description, interpretation and implications for regional marker tephras. *J. Volcanol. Geotherm. Res.* **275,** 114–131 (2014).
91. Sampson, D. E. & Cameron, K. L. The geochemistry of the Inyo volcanic chain—multiple magma systems in the Long Valley region, eastern California. *J. Geophys. Res.* **92,** 10403–10421 (1987).
92. Veres, D. *et al.* The Antarctic ice core chronology (AICC2012): an optimized multi-parameter and multi-site dating approach for the last 120 thousand years. *Clim. Past* **9,** 1733–1748 (2013).
93. Siebert, L., Simkin, T. & Kimberly, P. *Volcanoes of the World* 3rd edn, (University of California Press, 2010).



b Ice core records

Region	Name	Time Period	Depth (m)	^{10}Be analyses (N)	Latitude/Longitude	Accumulation (m water equ. a^{-1})	Reference
NW Greenland	NEEM-2011-S1	86-2006 CE	0-411	70	77.45°N, 51.06°W	0.210	ref. 16
	NEEM	500 BCE-1258 CE	183-514		77.45°N, 51.06°W	0.210	this study
NE Greenland	TUNU2013	280-2013 CE	0-212	9	78.04°N, 33.88°W	0.100	this study
Central Greenland	NGRIP	773-800 CE	245-250	27	75.12°N, 42.32°W	0.190	this study
W Antarctica	WDC	394 BCE-2006 CE	0-577	42	79.47°S, 112.09°W	0.210	ref. 16
E Antarctica	B40	500 BCE-2012 CE	0-200		75.00°S, 00.06°W	0.068	ref. 17, this study

Multi-centennial (>1,500 years) tree-ring chronologies at sites where temperature is the main limiting factor

Region	Name	Time Period	Parameter	Species	Latitude/Longitude	$R_{1901-2000}$ [BEST T_{JJA}]^{82}	Reference
Northern Finland	NScan	138 BCE-2006 CE	MXD	*Pinus sylvestris*	68°N, 23°E	0.71 ($p<0.0001$)†	ref. 43
Northern Sweden	Torneträsk	500-2000 CE	MXD	*Pinus sylvestris*	68°N, 19.5°E	0.71 ($p<0.0001$)†	ref. 77
Northern Siberia	Yamal*	1-2005 CE	RW	*Larix sibirica*	67°N, 70°E	0.45 ($p<0.0001$)†	ref. 76
Central Europe	Alps*	500 BCE-2009 CE	RW	*Pinus cembra, Larix decidua*	47°N, 11°E	0.69 ($p<0.0001$)†	ref. 42
Great Basin, USA	GB3rwi-IA2500*	500 BCE-2009 CE	RW	*Pinus longaeva*	38°N, 117°W	0.31 ($p=0.0011$)†	ref. 78
"Composite"	"N-Tree" (N=5)	500 BCE-2000 CE	MXD+RW	various	>38°N		this study

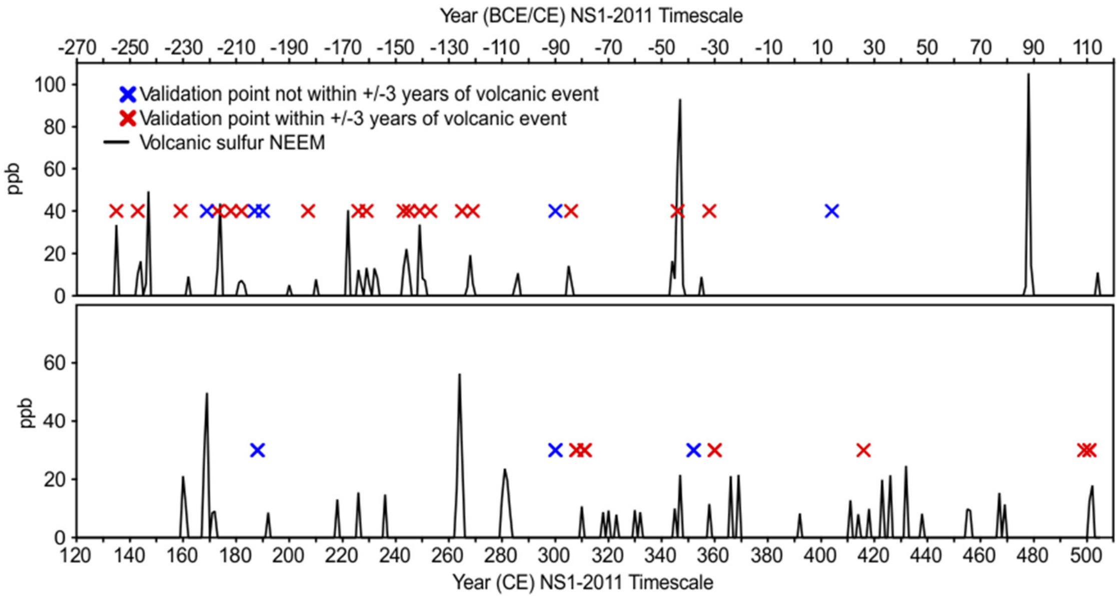
Summer temperature reconstructions (T_{JJA}; °C rel. 1961-1990)

Region	Data used	Time Period	Parameter	Species	$R_{1901-2000}$ [BEST T_{JJA}]^{82}	$R_{1900-2000}$ [HadCRUT4 T_{annual}]	Reference
N-Scandinavia	NScan + Torneträsk	1-2000 CE	MXD	*Pinus sylvestris*	0.74 ($p<0.0001$)†		ref. 35
Central Europe	Alps	1-2000 CE	RW	*Pinus cembra, Larix decidua*	0.69 ($p<0.0001$)†		ref. 42
Northern Siberia	Yamal	1-2005 CE	RW	*Larix sibirica*	0.45 ($p<0.0001$)†		ref. 76
Europe	NScan + Torneträsk + Alps + 8 tree-ring chronologies (<1,500 a)	1-2000 CE	MXD+RW	various		0.57 ($p<0.0001$)†	ref. 3

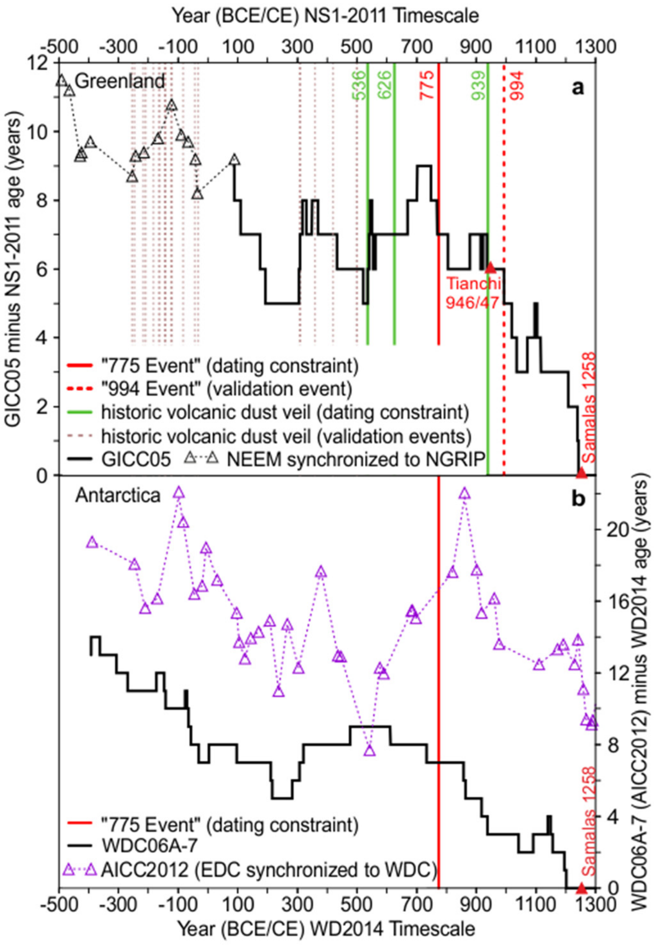
* the 775 event has been previously identified by radiocarbon measurements (ref. 12, 25)

† p-values are calculated using a (one-sided) Student's t-test, corrected for reduced sample sizes due to autocorrelation of the instrumental and proxy data

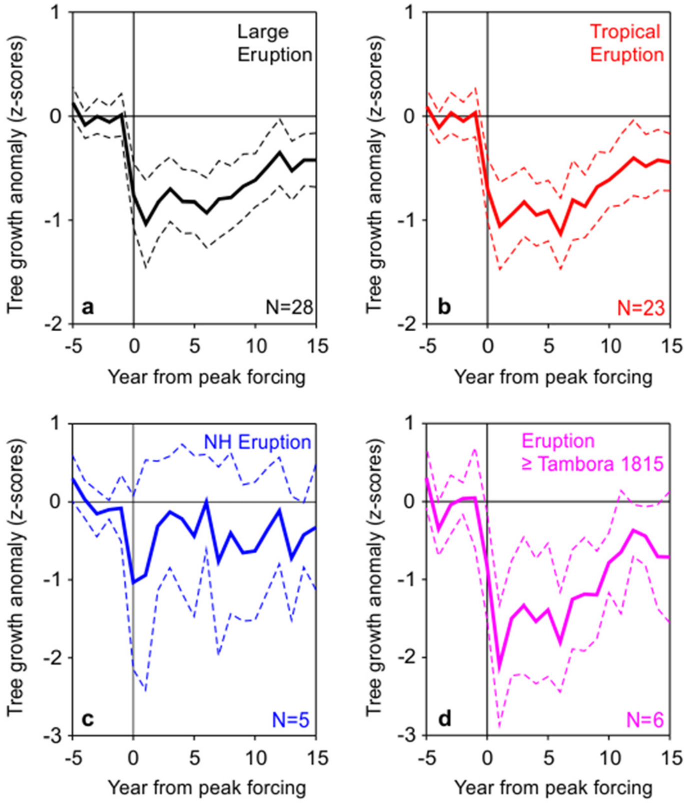
Extended Data Figure 1 | Location of study sites. a, Map showing locations (blue circles) of the five ice cores (WDC, B40, NEEM, NGRIP and TUNU) used in this study. Sites of temperature-limited tree-ring chronologies (green)[42,43,76–78] and sites with annual Δ^{14}C measurements from tree-rings in the eighth century CE (red outline) are marked. **b**, Metadata for the ice cores, tree-ring width (RW), maximum latewood density (MXD) chronologies and temperature reconstructions used[3,12,16,17,25,35,42,43,76,77,78,82]. m water equ. a^{-1}, metres of water equivalent per year.



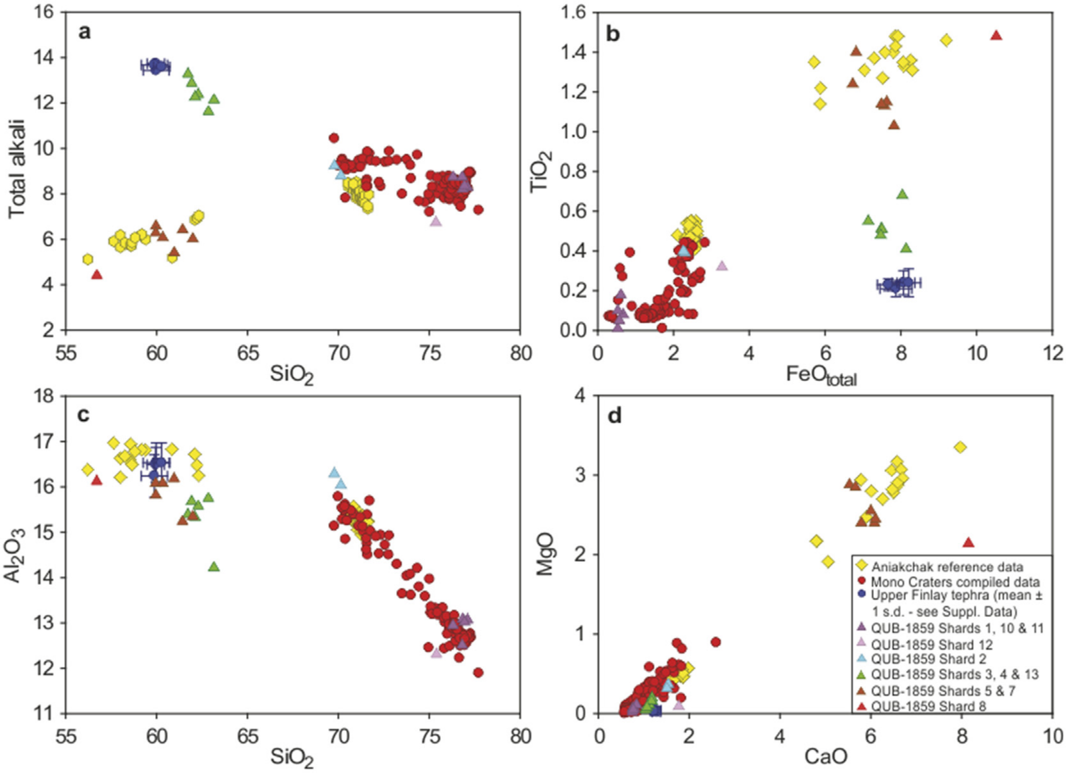
Extended Data Figure 2 | Volcanic dust veils from historical documentary sources in relation to NEEM. Time series of 32 independently selected chronological validation points from well dated historical observations of atmospheric phenomena with known association to explosive volcanism (for example, diminished sunlight, discoloured solar disk, solar corona or Bishop's Ring, red volcanic sunset) as reported in the Near East, Mediterranean region, and China, before our earliest chronological age marker at 536 CE. Black lines represent the magnitude (scale on *y* axes) of annual sulfate deposition measured in NEEM (NEEM and NEEM-2011-S1 ice cores) from explosive volcanic events on the new NS1-2011 timescale. Red crosses depict the 24 (75%) historical validation points for which NEEM volcanic events occur within a conservative ±3-year uncertainty margin. Blue crosses represent the eight points for which volcanic events are not observed. The association between validation points and volcanic events is statistically significantly non-random at>99.9% confidence ($P < 0.001$). ppb, parts per billion.



Extended Data Figure 3 | Timescale comparison. Age differences of the timescales NS1-2011 and GICC05 for the NEEM-2011-S1/NEEM ice cores (**a**) and WD2014 and WDC06A-7 for WDC (**b**). Differences before 86 CE (the age of the ice that is now at the bottom of the ice core NEEM-2011-S1) deriving from the annual-layer counting of the NEEM core are shown for major volcanic eruptions relative to the respective signals in NGRIP on the annual-layer counted GICC05 timescale. Marker events used for constraining the annual-layer dating (solid line) and for chronology evaluation (dashed lines) are indicated. Triangles mark volcanic signals. Also indicated is the difference between WD2014 and the Antarctic ice-core chronology (AICC2012)[92], based on volcanic synchronization between the WDC and EDC96 ice cores.



Extended Data Figure 4 | Post-volcanic suppression of tree growth. Superposed epoch analysis for large volcanic eruptions using the 28 largest volcanic eruptions (**a**); the 23 largest tropical eruptions (**b**); the five largest Northern Hemisphere eruptions (**c**); and eruptions larger than Tambora 1815 with respect to sulfate aerosol loading (**d**). Shown are growth anomalies of a multi-centennial tree-ring composite record (N-Tree) 15 years after the year of volcanic sulfate deposition, relative to the average of five years before the events. Dashed lines indicate 95% confidence intervals (2 s.e.m.) of the tree-ring growth anomalies associated with the multiple eruptions.



Extended Data Figure 5 | Major-element composition for ice core tephra QUB-1859 and reference material. Shown are selected geochemistry data: SiO$_2$ versus total alkali (K$_2$O + Na$_2$O) (**a**); FeO (total iron oxides) versus TiO$_2$ (**b**); SiO$_2$ versus Al$_2$O$_3$ (**c**); and CaO versus MgO (**d**) from 11 shards extracted from the NEEM-2011-S1 ice core at 327.17–327.25 m depth, representing the age range 536.0–536.4 CE on the new, NS1-2011 timescale. Data for Late Holocene tephra from Mono Craters (California) are from the compilation by ref. 90; data for Aniakchak (Alaska) are from reference material published by ref. 88; and data for the early Holocene upper Finlay tephra, believed to be from the Edziza complex in the Upper Cordilleran Volcanic province (British Columbia), are from ref. 89. (See Supplementary Information for the Upper Finlay tephra.)

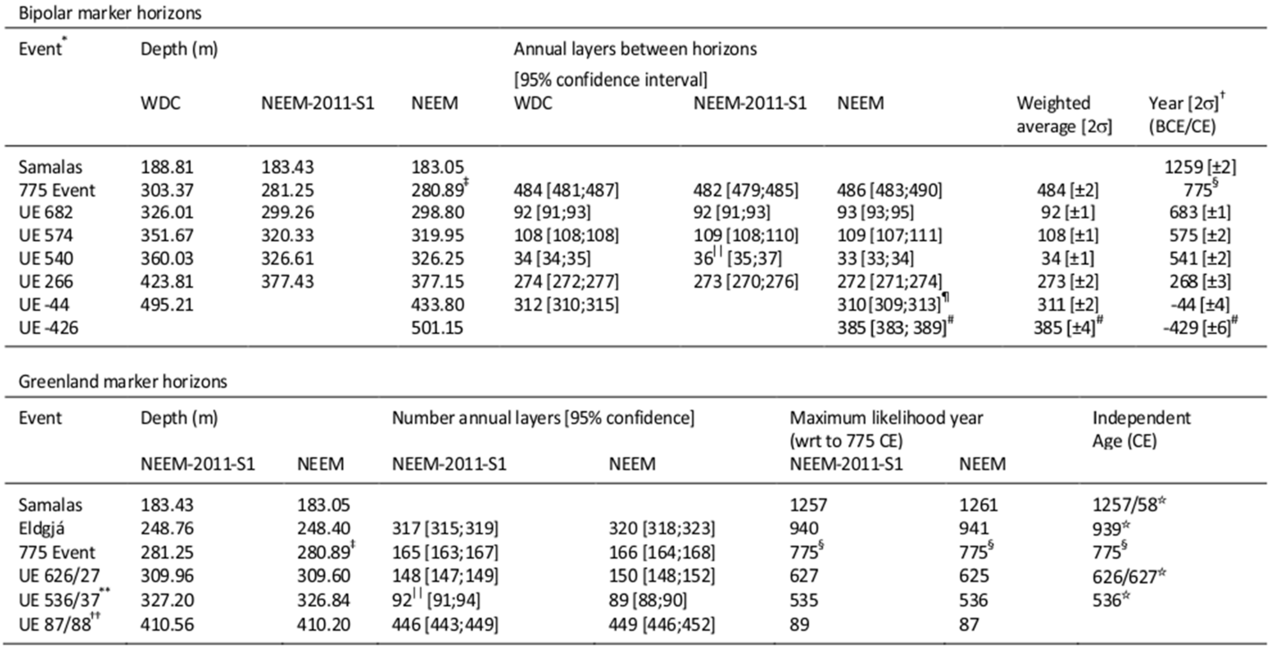
Extended Data Table 1 | Ice-core dating

Ice Core	Ice-core parameter	Dominant aerosol source	Deposition maximum
NEEM-2011-S1, (Greenland), 183-411m	Na, Cl nssS, nssS/Na nssCa, Mn, Ce BC, NH_4^+, BC geometric mean particle size Sr, Mg, I, NO_3^-	sea salt marine biogenic emissions dust biomass burning various	winter summer spring summer
NEEM, (Greenland), 183-410m	Na^+ Ca^{2+}, particle count NH_4^+, NO_3^- conductivity, ECM, H_2O_2	sea salt dust biomass burning various	winter spring summer
NEEM, (Greenland), 410-514m	Na, Sr, Cl, Na^+ nssS, nssS/Na nssCa, Mn, Mg, particle count, Ca^{2+} BC, NH_4^+, NO_3^- ,NH_4^+, NO_3^- conductivity, ECM, H_2O_2	sea salt marine biogenic emissions dust biomass burning various	winter summer spring summer
WDC, (Antarctica), 188-577m	Na, Sr nssS, nssS/Na, Br BC	sea salt marine biogenic emissions biomass burning	austral winter austral summer austral summer/fall

Year CE	Depth (m)	Event	Parameter	Independent age information
NEEM-2011-S1				
1258*	183.49	volcano (tropical)	nssS	ice core (GICC05, Law Dome)
994†	237.89	cosmic ray anomaly	^{10}Be	tree ring
946†	247.21	volcano (NH, Tianchi)	nssS, tephra	historical observation
939*	248.87	volcano (NH, Eldgjá)	nssS	historical observation
775*	281.45	cosmic ray anomaly	^{10}Be	tree ring
626*	310.01	volcano (NH)	nssS	historical observation
536*	327.23	volcano (NH)	nssS, tephra	historical observation
WDC				
1258*	188.91	volcano (tropical)	nssS	ice core (GICC05, Law Dome)
775*	303.36	cosmic ray anomaly	^{10}Be	tree ring

Parameters used for annual-layer interpretation. Parameters measured by the CFA system in the field are underlined. NH, Northern Hemisphere.
*Stratigraphic age marker used to constrain annual-layer counting.
†Horizons used to evaluate the timescale.

Extended Data Table 2 | Annual-layer results using the StratiCounter program



Bipolar marker horizons

Event[*]	Depth (m)			Annual layers between horizons [95% confidence interval]				
	WDC	NEEM-2011-S1	NEEM	WDC	NEEM-2011-S1	NEEM	Weighted average [2σ]	Year [2σ][†] (BCE/CE)
Samalas	188.81	183.43	183.05					1259 [±2]
775 Event	303.37	281.25	280.89[‡]	484 [481;487]	482 [479;485]	486 [483;490]	484 [±2]	775[§]
UE 682	326.01	299.26	298.80	92 [91;93]	92 [91;93]	93 [93;95]	92 [±1]	683 [±1]
UE 574	351.67	320.33	319.95	108 [108;108]	109 [108;110]	109 [107;111]	108 [±1]	575 [±2]
UE 540	360.03	326.61	326.25	34 [34;35]	36[||] [35;37]	33 [33;34]	34 [±1]	541 [±2]
UE 266	423.81	377.43	377.15	274 [272;277]	273 [270;276]	272 [271;274]	273 [±2]	268 [±3]
UE -44	495.21		433.80	312 [310;315]		310[309;313][¶]	311 [±2]	-44 [±4]
UE -426			501.15			385 [383; 389][#]	385 [±4][#]	-429 [±6][#]

Greenland marker horizons

Event	Depth (m)		Number annual layers [95% confidence]		Maximum likelihood year (wrt to 775 CE)		Independent Age (CE)
	NEEM-2011-S1	NEEM	NEEM-2011-S1	NEEM	NEEM-2011-S1	NEEM	
Samalas	183.43	183.05			1257	1261	1257/58[☆]
Eldgjá	248.76	248.40	317 [315;319]	320 [318;323]	940	941	939[☆]
775 Event	281.25	280.89[‡]	165 [163;167]	166 [164;168]	775[§]	775[§]	775[§]
UE 626/27	309.96	309.60	148 [147;149]	150 [148;152]	627	625	626/627[☆]
UE 536/37[**]	327.20	326.84	92[||] [91;94]	89 [88;90]	535	536	536[☆]
UE 87/88[††]	410.56	410.20	446 [443;449]	449 [446;452]	89	87	

Maximum-likelihood number of annual layers and confidence intervals derived from annual-layer counting between distinctive marker horizons and corresponding ages relative to the 775 CE ^{10}Be event. wrt, with respect to.

[*] Unattributed events (UE) give volcanic signal and year of sulfate deposition based on final age models.

[†] Year calculated from the number of annual layers relative to the fixed age marker in 775 CE (negative numbers are years BCE).

[‡] Depth has been estimated from the average depth offset between NEEM-2011-S1 and NEEM.

[§] Fixed age marker based on the ^{10}Be maximum annual value.

[||] Section with 6-m gap in the NEEM 2011-S1 core DRI data (this section is not used for calculating average age).

[¶] This section is based on the NEEM field CFA data, since the DRI data does not cover the entire interval.

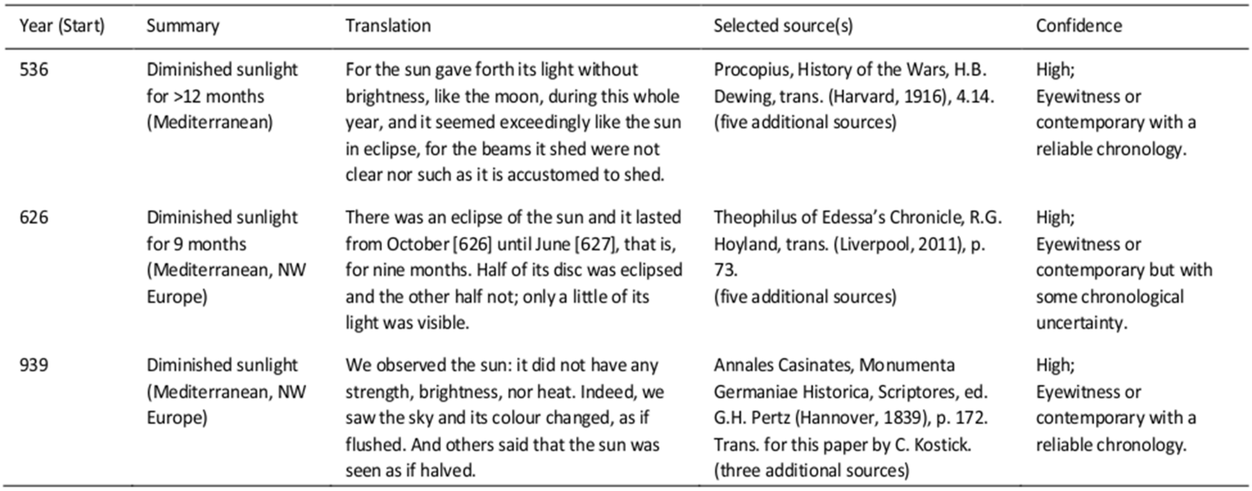
[#] Section is based on combined data set of DRI and field-measured CFA data. The number of annual layers in this section from manual interpretation by investigator 1 was 383 (±7), and that of investigator 2 was 393 (±8) layers. Most of the difference between the three layer counts occurred below 480 m (before 300 BCE), where data gaps were more frequent.

[☆] Independent age markers used to constrain annual-layer dating in a second iteration to derive the final ice-core age model NS1-2011.

[**] Tephra particles were extracted from the depth range 327.17–327.25 m (see Supplementary Data).

[††] Unattributed volcanic signal that was previously attributed to the historic 79 CE eruption of Vesuvius[64].

Extended Data Table 3 | Historical documentary evidence for key volcanic eruption age markers 536-939 CE



Year (Start)	Summary	Translation	Selected source(s)	Confidence
536	Diminished sunlight for >12 months (Mediterranean)	For the sun gave forth its light without brightness, like the moon, during this whole year, and it seemed exceedingly like the sun in eclipse, for the beams it shed were not clear nor such as it is accustomed to shed.	Procopius, History of the Wars, H.B. Dewing, trans. (Harvard, 1916), 4.14. (five additional sources)	High; Eyewitness or contemporary with a reliable chronology.
626	Diminished sunlight for 9 months (Mediterranean, NW Europe)	There was an eclipse of the sun and it lasted from October [626] until June [627], that is, for nine months. Half of its disc was eclipsed and the other half not; only a little of its light was visible.	Theophilus of Edessa's Chronicle, R.G. Hoyland, trans. (Liverpool, 2011), p. 73. (five additional sources)	High; Eyewitness or contemporary but with some chronological uncertainty.
939	Diminished sunlight (Mediterranean, NW Europe)	We observed the sun: it did not have any strength, brightness, nor heat. Indeed, we saw the sky and its colour changed, as if flushed. And others said that the sun was seen as if halved.	Annales Casinates, Monumenta Germaniae Historica, Scriptores, ed. G.H. Pertz (Hannover, 1839), p. 172. Trans. for this paper by C. Kostick. (three additional sources)	High; Eyewitness or contemporary with a reliable chronology.

A comprehensive list of sources, including translations and assessment of the confidence placed in each source and its chronological information is given in Supplementary Data. NW, northwest.

Extended Data Table 4 | Large volcanic eruptions during the past 2,500 years

Rank	Year	Volc. SO_4^{2-} Greenland (kg km^{-2})	Volc. SO_4^{2-} Antarctica (kg km^{-2})	Global forcing* (W m^{-2})	Cold year	N-Tree (z scores; 1000-99)	$T_{Europe/Arctic}$ (^{0}C; 1961-90)	Volcano†/ Region
1	-426	99.8	78.2	-35.6	-425	-2.74		UE -426
2	1258	90.4	73.4	-32.8	1258	-1.43	-0.91	Samalas/Indonesia
3	-44	100.6	15.4	-23.2	-43	-3.33		Chiltepe?/Nicaragua
4	1458	39.0	63.6	-20.5	1459	-2.31	-1.03	Kuwae/Vanuatu
5	540	61.2	34.4	-19.1	541	-2.57	-1.48	Ilopango?/El Salvador
6	1815	39.7	45.8	-17.1	1816	-2.51	-1.55	Tambora/Indonesia
7	1230	56.4	23.1	-15.9	1230	-1.71	-0.65	UE 1230
8	1783	135.8		-15.5	1783	-1.16	-0.97	Laki/Iceland
9	682	38.4	38.7	-15.4	682	-0.95	-0.96	Pago?/New Britain
10	574	38.3	34.1	-14.5	574	-2.46	-0.94	Rabaul?/New Britain
11	266	61.0	11.3	-14.5	268	-1.70	-0.72	UE 266
12	1809	34.6	25.4	-12.0	1810	-2.18	-1.23	UE 1809
13	1108	48.3	11.6	-12.0	1109	-1.99	-1.15	UE 1108
14	1641	44.2	14.9	-11.8	1641	-2.31	-1.19	Parker/Philippines
15	1601	39.2	18.7	-11.6	1601	-2.62	-1.50	Huaynaputina/Peru
16	169	39.1	18.4	-11.5	170	-0.80	-0.94	UE 169
17	1171	37.0	19.5	-11.3	1171	-0.91	-0.88	UE 1171
18	536	99.0		-11.3	536	-3.36	-1.74	UE 536
19	1695	28.6	22.5	-10.2	1696	-1.63	-1.28	UE 1695
20	939	88.7		-10.1	940	-1.81	-1.44	Eldgjá/Iceland
21	1286	27.6	20.8	-9.7	1288	-1.49	-0.65	Quilotoa?/Ecuador
22	433	20.6	27.2	-9.6	432	-0.45	-0.25	UE 433
23	87	83.1		-9.5	87	-0.22	-0.49	UE 87
24	1345	27.9	19.1	-9.4	1346	-2.18	-1.48	El Chichon?/Mexico
25	626	72.2		-8.2	627	-3.00	-0.93	UE 626

Years with negative numbers are BCE. Tentative attribution of ice-core signals to historic volcanic eruptions is based on the Global Volcanism Program volcanic eruption database[93]. Average (summer) temperature for the associated cold year is given for the average of Europe and the Arctic[3]. Volc., volcanic.

*Total global aerosol forcing was estimated by scaling the total sulfate flux from both polar ice sheets to the reconstructed total (that is, time integrated) aerosol forcing for Tambora 1815[7] (Methods); for high-latitude Northern Hemisphere eruptions, Greenland fluxes were scaled by a factor of 0.57[6].

†Unattributed volcanic events (UE) and tentative attributions for non-documented historic eruptions (?) are marked.

Extended Data Table 5 | Post-volcanic cooling

Rank	Year	JJA temperature anomaly (°C)	Volcanic event(s)	Decade	JJA temperature anomaly (°C)	Volcanic event(s)
1	1821	-1.82	**1815**, 1821	1600-1609	-1.17	1595, **1601**
2	1601	-1.82	**1600**	536-545	-1.12	**536**, **540**
3	1675	-1.78	1673	1812-1821	-1.10	**1809**, **1815**
4	536	-1.67	536	1453-1462	-1.01	**1453**, **1458**
5	800	-1.66	800	1587-1596	-0.98	1585, 1590, **1595**
6	1816	-1.64	**1815**	1107-1116	-0.96	**1108**, 1115
7	1453	-1.57	**1453**	1344-1353	-0.87	1341, **1345**
8	1633	-1.56		351-360	-0.83	351, 358, 360
9	1109	-1.56	**1108**	1692-1701	-0.82	1693, **1695**
10	1608	-1.53		413-422	-0.79	411, 418
11	544	-1.50	**540**	1463-1472	-0.79	**1458**, 1463, 1470
12	574	-1.48	**574**	1127-1136	-0.79	1127
13	1695	-1.47	1693, **1695**	389-398	-0.78	388, 393
14	543	-1.46	**540**	1672-1681	-0.77	1673
15	541	-1.43	**540**	1632-1641	-0.76	1637, **1641**
16	549	-1.43	547	1258-1267	-0.76	**1258**

Rank	Year	Tree growth anomaly (z-scores)	Volcanic event(s)	Decade	Tree growth anomaly (z-scores)	Volcanic event(s)
1	536*	-3.4	**536**	536-545	-2.2	**536**, **540**
2	-43*	-3.3	**-44**	1812-1821	-2.2	**1809**, **1815**
3	627*	-3.0	**626**	1453-1462	-2.0	**1453**, **1458**
4	543	-3.0	**540**	-43 to -34	-2.0	-46, **-44**
5	-360	-2.9	-360	1601-1610	-2.0	**1600**
6	-35	-2.8	-35	-361 to -352	-1.8	-360,-356
7	-425*	-2.7	**-426**	1463-1472	-1.7	**1458**, 1463
8	-42*	-2.7	**-44**	1832-1841	-1.7	**1831**, **1835**
9	546	-2.6	**540**	1341-1350	-1.7	**1344**
10	-140*	-2.6	-141	546-555	-1.6	**540**, 547
11	541	-2.6	**540**	1673-1682	-1.5	1673
12	544	-2.6	**540**	-427 to -418	-1.5	**-426**
13	545	-2.5	**540**	1330-1339	-1.4	1329, 1336
14	574*	-2.5	**574**	1638-1647	-1.4	**1641**, 1646
15	-354*	-2.5	-356	1699-1708	-1.4	**1695**, 1708
16	-38	-2.5		1285-1294	-1.4	**1286**

Coldest years and decades (1–2000 CE, JJA temperature with respect to 1901–2000) for Europe[3] and years (500 BCE–1250 CE) and decades (500 BCE–2000 CE) with strong growth reduction in the N-Tree composite (with respect to 1000–1099). Ages of the volcanic events from the ice cores reflect the start of volcanic sulfate deposition in Greenland (NS1-2011 timescale) with the largest 40 events indicated in bold letters and tropical eruptions underlined. Years with negative numbers are before the Common Era (BCE).

*Latewood frost ring in bristlecone pines within one year[34].

doi:10.1038/nature14660

Metabolic co-dependence gives rise to collective oscillations within biofilms

Jintao Liu[1], Arthur Prindle[1]*, Jacqueline Humphries[1]*, Marçal Gabalda-Sagarra[2]*, Munehiro Asally[3]*, Dong-yeon D. Lee[1], San Ly[1], Jordi Garcia-Ojalvo[2] & Gürol M. Süel[1]

Cells that reside within a community can cooperate and also compete with each other for resources. It remains unclear how these opposing interactions are resolved at the population level. Here we investigate such an internal conflict within a microbial (*Bacillus subtilis*) biofilm community: cells in the biofilm periphery not only protect interior cells from external attack but also starve them through nutrient consumption. We discover that this conflict between protection and starvation is resolved through emergence of long-range metabolic co-dependence between peripheral and interior cells. As a result, biofilm growth halts periodically, increasing nutrient availability for the sheltered interior cells. We show that this collective oscillation in biofilm growth benefits the community in the event of a chemical attack. These findings indicate that oscillations support population-level conflict resolution by coordinating competing metabolic demands in space and time, suggesting new strategies to control biofilm growth.

Cooperation and competition are complex social interactions that can have critical roles in biological communities. Cooperative behaviour often increases the overall fitness of the population through processes such as division of labour and production of common goods[1–4]. At the same time, individuals in a community compete with each other for limited resources, such as nutrients[5,6]. Here we investigated bacterial biofilms[7–10] to determine how the conflict between the opposing social behaviours of cooperation and competition could be resolved at the community level to increase overall fitness.

Biofilms typically form under environmental stress conditions, such as nutrient limitation[11–13]. As these bacterial communities grow larger, the supply of nutrients to interior cells becomes limited due to an increase in nutrient consumption associated with the growth of multiple layers of cells in the biofilm periphery. Severe nutrient limitation for interior cells is detrimental to the colony, since the sheltered interior cells are critical to the survival of the biofilm community in the event of an external challenge. This defines a fundamental conflict between the opposing demands for biofilm growth and maintaining the viability of protected (interior) cells (Fig. 1a). The identification of possible mechanisms that ensure the viability of the protected interior cells is fundamental to understanding biofilm development[14,15].

To investigate directly how *Bacillus subtilis* biofilms continue expanding while sustaining interior cells, we converted the potentially complex three-dimensional problem to a simpler two-dimensional scenario using microfluidics. Specifically, we used growth chambers that are unconventionally large in the lateral, *x*–*y* dimensions (3 × 3 mm), while confining biofilm thickness (*z* dimension) to only a few micrometres (Fig. 1b). Therefore, biofilm expansion in this device is predominantly limited to two dimensions, creating a ‘pancake-like’ configuration. In fact, biofilms often form in confined aqueous environments and thus this microfluidic chamber may better mimic those growth conditions[11–13]. This experimental set-up is thus ideal to interrogate how biofilms can reconcile the opposing benefits of growth and protection during biofilm development.

Oscillations in biofilm growth

Unexpectedly, we observed oscillations in biofilm expansion despite constant media flow within the microfluidic device (Fig. 1c, d, Supplementary Video 1 and Extended Data Fig. 1a). Specifically, biofilms exhibit periodic reduction in colony expansion that is self-sustained and can last for more than a day (Fig. 1e and Extended Data Fig. 1b). The period of oscillations has a mean of 2.5 ± 0.8 h (standard deviation (s.d.), $n = 63$ colonies), which is less than the duration of the average cell replication time of 3.4 ± 0.2 h (s.d., $n = 21$ cell cycles) under this growth condition (Fig. 1f and Methods, ‘Data analysis’ section). Moreover, oscillations only arise when the biofilm exceeds a certain colony size (Supplementary Video 2). In particular, quantitative measurements obtained from 53 individual biofilms indicate that oscillations emerge in colonies that exceed an average diameter of 580 ± 85 μm (s.d., $n = 53$ colonies), which corresponds to approximately one million cells (Fig. 1g, h). Together, these data show that oscillations arise during biofilm formation and are self-sustained.

Given that biofilms typically form under nutrient-limited conditions and bacterial growth is generally controlled by metabolism, we hypothesized that metabolic limitation has a key role in the observed periodic halting of biofilm expansion. In particular, after determining that carbon source limitation did not have an essential role in the oscillations (Extended Data Fig. 2), we focused on nitrogen limitation. The standard biofilm growth media (MSgg, see Methods, ‘Growth conditions’ section) used to study *B. subtilis* biofilm development contains glutamate as the only nitrogen source[16]. In most organisms including *B. subtilis*, glutamate is combined with ammonium by glutamine synthetase (GS) to produce glutamine, which is essential for biomass production and growth (Fig. 2a)[17]. Cells can obtain the necessary ammonium from glutamate through the enzymatic activity of glutamate dehydrogenase (GDH), expressed by the *rocG* or *gudB* genes in the undomesticated *B. subtilis* used in this study (Fig. 2a)[18–20]. To determine whether biofilms experience glutamine limitation, we measured expression of *nasA*, one of several genes activated in

[1]Division of Biological Sciences, University of California San Diego, California 92093, USA. [2]Department of Experimental and Health Sciences, Universitat Pompeu Fabra, 08003 Barcelona, Spain. [3]Warwick Integrative Synthetic Biology Centre, School of Life Sciences, University of Warwick, Coventry CV4 7AL, UK.
*These authors contributed equally to this work.

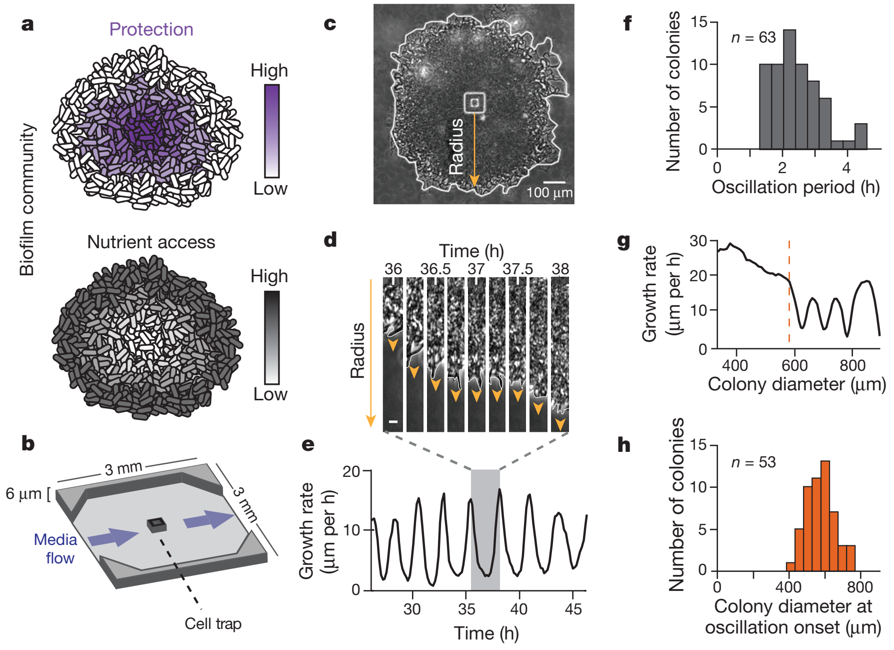


Figure 1 | Biofilms grown in microfluidic devices show oscillations in colony expansion. **a**, Biofilms must reconcile opposing demands for protection from external challenges (gradient indicated in purple) and access to nutrients (gradient indicated in grey). **b**, Schematic of the microfluidic device used throughout this study. Direction of media flow is indicated by the blue arrows. **c**, Phase contrast image of a biofilm growing in the microfluidic device. The yellow arrow indicates the region of interest in panel **d**. **d**, Film strip of a radius of the biofilm over time shows a pause in colony expansion. This film strip represents one cycle of biofilm oscillations, indicated by the shaded region in panel **e**. Scale bar, 5 µm. The arrowheads indicate direction of biofilm growth. **e**, Growth rate over time shows persistent oscillations in colony expansion. **f**, Histogram of the average period of oscillations for each colony (*n* = 63 colonies, mean = 2.5 h, s.d. = 0.8 h). The cell replication time is approximately 3.4 h under these conditions (Methods, 'Data analysis' section). **g**. Growth rate as a function of colony diameter (which increases in time) shows that early colony growth does not exhibit oscillations. The orange line indicates the diameter (~600 µm) at which this colony initiates oscillations. **h**, Histogram of the diameter at which a colony begins to oscillate (*n* = 53 colonies, mean = 576 µm, s.d. = 85 µm).

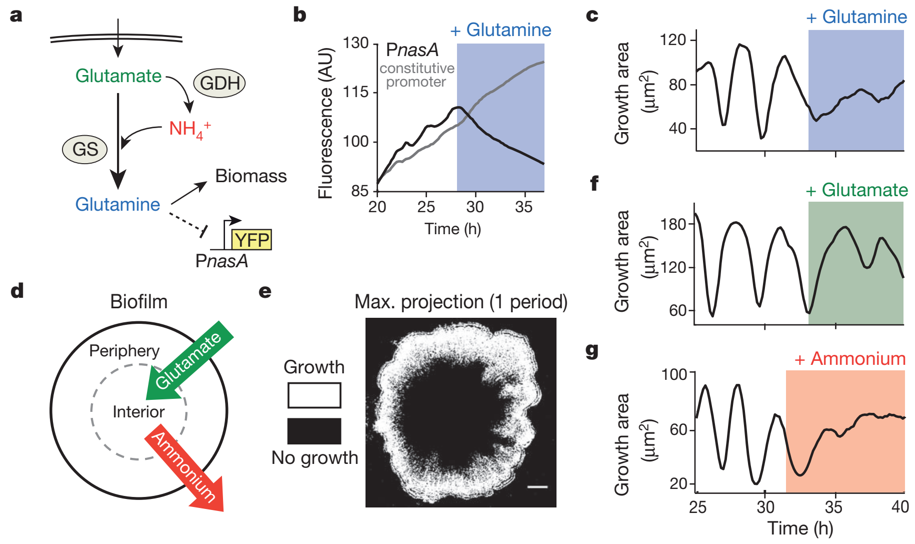


Figure 2 | Biofilm growth depends specifically on extracellular ammonium availability. **a**, Colony growth in MSgg medium depends on the production of glutamine from externally supplied glutamate and self-produced or scavenged ammonium. Glutamine limitation was monitored using yellow fluorescent protein (YFP) expressed from the *nasA* promoter, which is activated upon glutamine limitation[21]. **b**, Addition of 1 mM glutamine (blue shading) represses expression from the P*nasA*-YFP reporter (black), but does not affect expression from a constitutive reporter (P*hyperspank*-CFP + 1 mM IPTG, grey). **c**, Growth area (see Methods, 'Data analysis' section) before and after addition of 1 mM glutamine to an oscillating colony. **d**, Of the two nutrients required for glutamine production, externally supplied glutamate (green) is most abundant in the biofilm periphery, while biofilm-produced ammonium (red) is most abundant in the biofilm interior. **e**, Maximum intensity projection over one period of a colony oscillation, made from a difference movie (Methods, 'Data analysis' section), which shows regions of growth (white) and no growth (black). Scale bar, 100 µm. **f**, Growth area of an oscillating colony before and after addition of 30 mM glutamate (green shading). **g**, Growth area of an oscillating colony before and after addition of 1 mM ammonium (red shading).

response to a lack of glutamine[21]. Results show that biofilms indeed experience glutamine limitation during growth. Specifically, supplementation of growth media directly with glutamine reduced *nasA* promoter expression, but did not affect expression of a constitutive promoter, confirming glutamine limitation within the biofilm (Fig. 2b). More strikingly, addition of exogenous glutamine eliminated periodic halting of biofilm growth (Fig. 2c and Extended Data Fig. 3a). These findings suggest that glutamine limitation plays a critical part in the observed oscillations during biofilm expansion.

The synthesis of glutamine requires both glutamate and ammonium; therefore, we investigated which of these substrates could be responsible for the observed glutamine limitation. Glutamate is provided in the media and is thus readily available to cells in the periphery of the biofilm. However, consumption of glutamate by peripheral cells is likely to limit its availability to cells in the biofilm interior (Fig. 2d). One may thus expect that oscillations in biofilm expansion could be due to periodic pausing of cell growth in the biofilm interior. Accordingly, we set out to establish whether interior or peripheral cells exhibited changes in growth. By tracking physical movement within the biofilm, we uncovered that only peripheral cells grow, and that oscillations in biofilm expansion therefore arise exclusively from periodic halting of peripheral cell growth (Fig. 2e, Supplementary Video 3, Extended Data Fig. 4a and Methods, 'Data analysis' section). This finding was further confirmed by single-cell resolution analysis that directly showed periodic reduction in the growth of peripheral cells (Extended Data Fig. 4b). This surprising pausing of cell growth in the periphery, despite unrestricted access to glutamate, suggests that glutamate cannot be the limiting substrate for glutamine synthesis. Consistent with this expectation, biofilm oscillations were not quenched by supplementation of the media with glutamate (Fig. 2f). Therefore, it is not glutamate but ammonium that appears to be the limiting substrate for glutamine synthesis in the biofilm periphery.

Because cells can self-produce ammonium from glutamate, we next sought to determine how peripheral cells could experience periodic ammonium limitation despite a constant supply of glutamate in the media. It is well known that ammonium production is a highly regulated process that is dependent on the metabolic state of the cell and the ambient level of ammonium in the environment[22]. In particular, since ammonium is in equilibrium with ammonia vapour, which can freely cross the cell membrane and be lost to the extracellular media[23], the production of ammonium is known as a 'futile cycle'. Cells therefore preferentially use extracellular (ambient) ammonium for growth, rather than producing their own[24–26]. Since peripheral cells are exposed to media flow, they are particularly susceptible to this futile cycle of ammonia loss. In this sense, as ammonium is not provided in the media, even if all cells produce ammonium, the biofilm interior will be the major source for ambient ammonium (Fig. 2d). Consequently, the simplifying hypothesis is that growth of peripheral cells relies on ammonium produced within the biofilm. To test this conjecture, we supplemented the media with 1 mM ammonium, which eliminated the periodic halting in biofilm expansion (Fig. 2g and Extended Data Figs 3b and 5a). When additional ammonium was suddenly removed from the media, growth in the biofilm periphery halted, as expected (Extended Data Fig. 5b). These findings indicate that peripheral cells preferentially rely on extracellular ammonium produced within the biofilm for their growth.

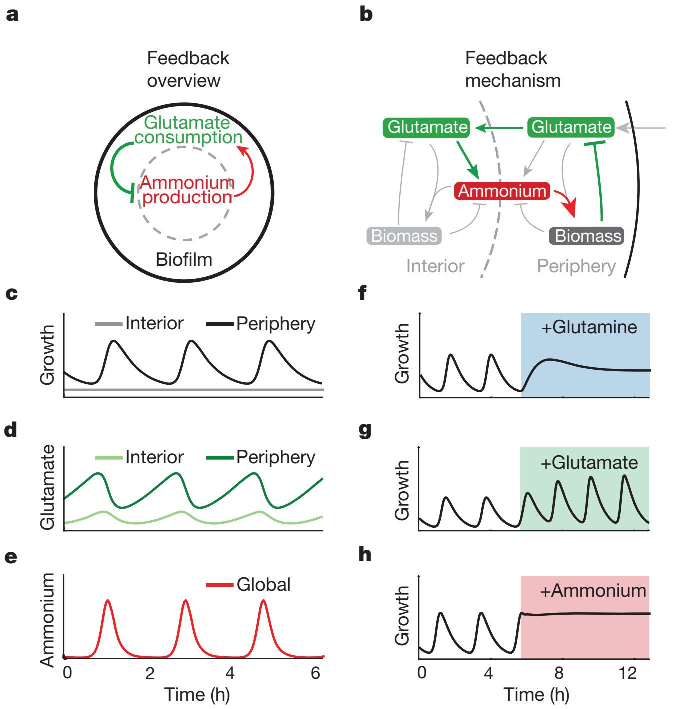


Figure 3 | Mathematical modelling of a spatial metabolic feedback loop gives rise to oscillations consistent with experimental data. a, The production of ammonium in the interior is limited by and at the same time triggers the consumption of glutamate in the periphery (green and red arrows, respectively), producing a delayed negative feedback loop. **b**, The excess glutamate not consumed by the biofilm periphery diffuses to the interior, where it can be converted into ammonium (green arrows). The ammonium in turn enhances growth in the periphery (red arrow) and consequently reduces the supply of glutamate to the interior. Model predictions are shown in **c**–**h**. **c**, Biofilm growth over time. **d**, Glutamate concentration over time. **e**, Ammonium concentration over time. **f**, Colony growth before and after glutamine addition (indicated by blue shading). **g**, Colony growth before and after addition of glutamate (green shading). **h**, Colony growth before and after addition of ammonium (red shading).

Metabolic co-dependence within the biofilm

The results described above evoke the possibility that ammonium limitation for peripheral cells may arise due to glutamate limitation for interior cells. Specifically, persistent consumption of glutamate by peripheral cells can deprive the interior cells of the necessary glutamate for ammonium production. To explore this nontrivial hypothesis, we turned to mathematical modelling to develop a conceptual framework and generate experimentally testable predictions. Our model describes separately the metabolic dynamics of interior and peripheral cells and the metabolite exchange between them, where the distinction of the two subpopulations depends on nutrient availability (see Supplementary Information, 'Mathematical Model' section). The model thus consists of two main assumptions (Fig. 3a). First, consumption of glutamate during growth of peripheral cells deprives interior cells of this nutrient and thus inhibits ammonium production in the biofilm interior. Second, the growth of peripheral cells depends predominantly on ammonium that is produced by metabolically stressed interior cells. A model based on these two simplifying assumptions (Fig. 3b) generates oscillations consistent with our experimental observations (Fig. 3c–e) and reproduces the effects of supplementing the media with glutamine, glutamate and ammonium (Fig. 3f–h, Extended Data Fig. 6 and Supplementary Information, 'Mathematical Model' section). The model also accounts for the observed slight increase of the oscillation period by considering an increase in the ratio of interior to peripheral cells over time (Extended Data Figs 1b and 6f). Therefore, this simple model shows that periodic halting in biofilm growth can result from metabolic co-dependence between cells in the biofilm periphery and interior that is driven by glutamate consumption and ammonium production, respectively.

The metabolic co-dependence between interior and peripheral cells gives rise to the surprising prediction that external attack could promote growth within the biofilm. Specifically, killing of peripheral cells will eliminate their glutamate consumption, which will increase glutamate availability in the biofilm and thereby promote growth of

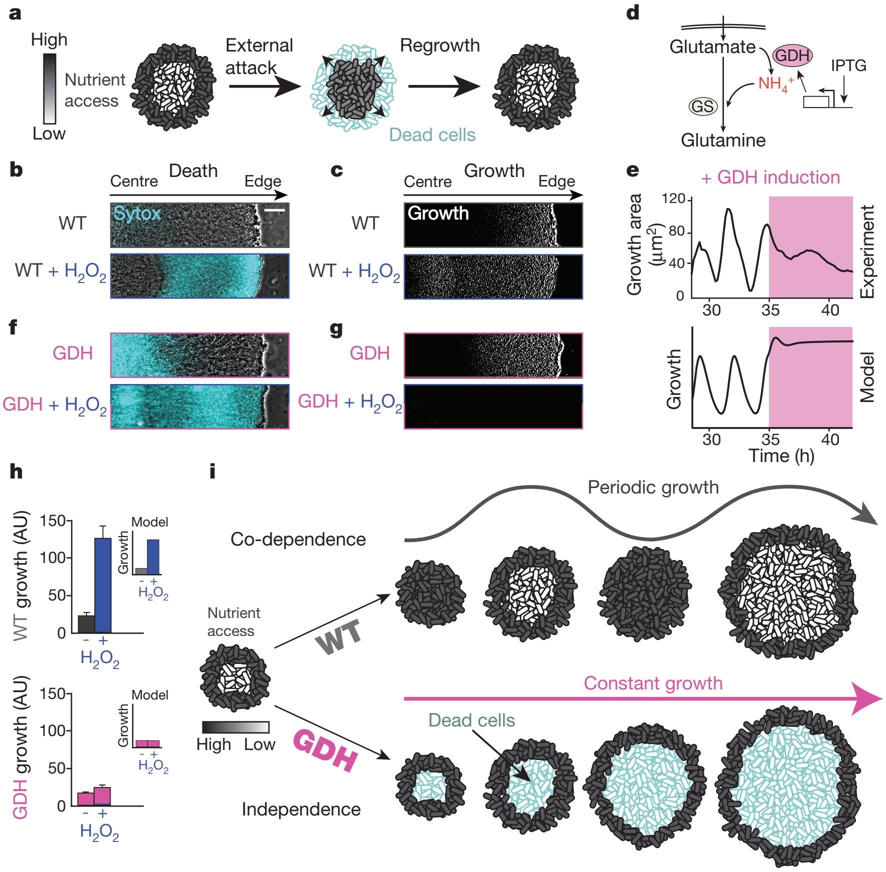


Figure 4 | Metabolic co-dependence between interior and peripheral cells gives rise to oscillations that make the colony more resilient to external attack. **a**, Visual representation of the predicted outcome of an external attack on biofilm growth. **b**, Phase contrast merged with cell death marker (cyan, 1 μM Sytox green) images of a wild-type (WT) biofilm region shows cell death with and without challenge by 2% (v/v) H_2O_2. Scale bar, 50 μm. **c**, In the same biofilm, difference images (white regions indicate cell growth) show wild-type growth with and without challenge by H_2O_2. **d**, Overexpression of glutamate dehydrogenase (GDH, pink) promotes more production of ammonium from glutamate. **e**, Experimental (top) and modelling results (bottom) of GDH overexpression (induced with 1 mM IPTG, indicated by pink shading). **f**, Phase contrast merged with cell death marker (cyan, 1 μM Sytox green) images of a colony overexpressing GDH with and without challenge by H_2O_2. **g**, In the same biofilm, difference images show cell growth during GDH overexpression alone, and with challenge by H_2O_2. **h**, Quantification of total biofilm growth rate in wild-type (top, $n = 4$ colonies) and GDH overexpression (bottom, $n = 3$ colonies) strains upon challenge with H_2O_2. Error bars represent standard deviations. Modelling data are shown as an inset for each strain.
i, Co-dependence between interior and peripheral cells exhibited in a wild-type strain results in a growth strategy that sustains the viability of interior cells, while independence enforced by a GDH overexpression strain results in starvation of interior cells and reduced resilience to external attack.

interior cells (Fig. 4a). To test this hypothesis, we measured cell death and growth within oscillating biofilms (Fig. 4b, top, and Extended Data Fig. 7). When we exposed the biofilm to media containing hydrogen peroxide (H_2O_2), we observed increased cell death predominantly in the biofilm periphery (Fig. 4b, bottom, and Extended Data Fig. 8). As predicted, death of peripheral cells led to growth of interior cells (Fig. 4c and Extended Data Fig. 8). To verify that this response is not uniquely triggered by H_2O_2, we exposed biofilms to the antibiotic chloramphenicol and again observed growth of interior cells (Extended Data Fig. 8). These findings further support our hypothesis that glutamate consumption by peripheral cells limits its availability in the biofilm.

The benefit of biofilm oscillations

Our model also assumes that glutamate starvation of the biofilm interior reduces the production of ammonium that can support peripheral cell growth. This assumption provokes the question as to why peripheral cells do not simply overcome their dependence on extracellular ammonium by increasing intracellular production[27,28]. To address this question, we constructed a strain that contains an inducible copy of the GDH gene *rocG* (Fig. 4d). We confirmed that GDH overexpression was not toxic to individual cells and did not affect their growth rate (Extended Data Fig. 9). In contrast, the induction of GDH expression in the biofilm quenched growth oscillations (Fig. 4e and Extended Data Fig. 3c) and resulted in high levels of cell death in the colony interior (Fig. 4f, top). This result explains why peripheral cells do not appear to utilize the simple strategy of overcoming their dependence on extracellular ammonium: such a strategy would result in the continuous growth of peripheral cells, starving and ultimately causing the death of sheltered interior cells within the biofilm. Periodic halting of peripheral cell growth due to extracellular ammonium limitation thus promotes the overall viability of the biofilm.

The ability of the biofilm to regenerate itself in the event of an external attack suggested that killing the biofilm interior first would be a more effective strategy for biofilm elimination. Accordingly, we exposed the GDH overexpression strain to hydrogen peroxide and again measured growth and death. As described above, GDH induction causes death of interior cells. Exposing the GDH overexpression strain to hydrogen peroxide resulted in more effective global killing throughout the biofilm (Fig. 4f, g, bottom). While in the wild-type biofilm, interior cells begin to grow in response to an external attack, metabolic independence between interior and peripheral cells in the GDH strain interferes with this defence mechanism (Fig. 4h). This outcome is also consistent with modelling predictions (Fig. 4h, inset). Oscillations in biofilm growth that are driven by metabolic co-dependence thus promote the resilience of the biofilm community by sustaining the viability of the sheltered interior cells that are most likely to survive in the event of an environmental stress (Fig. 4i).

Discussion

The data presented here reveal that intracellular metabolic activity within biofilms is organized in space and time, giving rise to co-dependence between interior and peripheral cells. Even though bacteria are single-celled organisms, the metabolic dynamics of individual cells can thus be regulated in the context of the community. This metabolic co-dependence can, in turn, give rise to collective oscillations that emerge during biofilm formation and promote the resilience of biofilms against chemical attack. The community-level oscillations also support the ability of biofilms to reach large sizes, while retaining a viable population of interior cells. Specifically, periodic halting of peripheral cell growth prevents complete starvation and death of the interior cells. This overcomes the colony size limitation for a viable biofilm interior that would otherwise be imposed by nutrient consumption in the biofilm periphery. Metabolic co-dependence in biofilms therefore offers an elegant solution that resolves the social conflict between cooperation (protection) and competition (starvation) through oscillations.

The discovery of biofilm oscillations presented here also raises new questions. While cellular processes such as swarming or expression of extracellular matrix components are not required for the observed biofilm oscillations (Extended Data Fig. 10), it will be interesting to pursue whether such cellular processes are influenced by oscillatory dynamics[29]. Another question worth pursuing is whether metabolic co-dependence can also arise in other biofilm-forming species. Perhaps other metabolic branches where metabolites can be shared among cells could also give rise to oscillations in biofilm growth. It will be interesting to pursue these questions in future studies to obtain a better understanding of biofilm development.

Our observations also suggest future strategies to cope with the intriguing resilience of biofilms in the face of environmental stresses, such as antibiotic exposure. In particular, our findings show that straightforward application of stress (such as H_2O_2 or chloramphenicol) to the biofilm counterintuitively promotes growth, effectively rejuvenating the biofilm. Death of the colony periphery relieves the repression on the growth of interior cells, allowing them to regenerate a new biofilm periphery and interior. In contrast, manipulation of the metabolic co-dependence may yield a more effective approach to control biofilm formation. Specifically, promoting continuous growth of peripheral cells can starve the biofilm interior, leaving behind the exposed peripheral cells that can more easily be targeted by external killing factors. Therefore, the metabolically driven collective oscillations in biofilm expansion described here not only reveal fundamental insights into the principles that govern formation of multicellular communities, but also suggest new strategies for manipulating the growth of biofilms.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 March; accepted 12 June 2015.

Published online 22 July 2015.

1. Ben-Jacob, E., Cohen, I. & Levine, H. Cooperative self-organization of microorganisms. *Adv. Phys.* **49,** 395–554 (2000).
2. Eldar, A. Social conflict drives the evolutionary divergence of quorum sensing. *Proc. Natl Acad. Sci. USA* **108,** 13635–13640 (2011).
3. Gregor, T., Fujimoto, K., Masaki, N. & Sawai, S. The onset of collective behavior in social amoebae. *Science* **328,** 1021–1025 (2010).
4. Wingreen, N. S. & Levin, S. A. Cooperation among microorganisms. *PLoS Biol.* **4,** 1486–1488 (2006).
5. Hibbing, M. E., Fuqua, C., Parsek, M. R. & Peterson, S. B. Bacterial competition: surviving and thriving in the microbial jungle. *Nature Rev. Microbiol.* **8,** 15–25 (2010).
6. Oliveira, N. M., Niehus, R. & Foster, K. R. Evolutionary limits to cooperation in microbial communities. *Proc. Natl Acad. Sci. USA* **111,** 17941–17946 (2014).
7. Davies, D. Understanding biofilm resistance to antibacterial agents. *Nature Rev. Drug Discov.* **2,** 114–122 (2003).
8. Donlan, R. M. & Costerton, J. W. Biofilms: survival mechanisms of clinically relevant microorganisms. *Clin. Microbiol. Rev.* **15,** 167–193 (2002).
9. Vlamakis, H., Aguilar, C., Losick, R. & Kolter, R. Control of cell fate by the formation of an architecturally complex bacterial community. *Genes Dev.* **22,** 945–953 (2008).
10. Yildiz, F. H. & Visick, K. L. *Vibrio* biofilms: so much the same yet so different. *Trends Microbiol.* **17,** 109–118 (2009).
11. Berk, V. *et al.* Molecular architecture and assembly principles of *Vibrio cholerae* biofilms. *Science* **337,** 236–239 (2012).
12. Costerton, J. W., Stewart, P. S. & Greenberg, E. P. Bacterial biofilms: a common cause of persistent infections. *Science* **284,** 1318–1322 (1999).
13. Hall-Stoodley, L., Costerton, J. W. & Stoodley, P. Bacterial biofilms: from the natural environment to infectious diseases. *Nature Rev. Microbiol.* **2,** 95–108 (2004).
14. Asally, M. *et al.* Localized cell death focuses mechanical forces during 3D patterning in a biofilm. *Proc. Natl Acad. Sci. USA* **109,** 18891–18896 (2012).
15. Wilking, J. N. *et al.* Liquid transport facilitated by channels in *Bacillus subtilis* biofilms. *Proc. Natl Acad. Sci. USA* **110,** 848–852 (2013).
16. Branda, S. S., Gonzalez-Pastor, J. E., Ben-Yehuda, S., Losick, R. & Kolter, R. Fruiting body formation by *Bacillus subtilis*. *Proc. Natl Acad. Sci. USA* **98,** 11621–11626 (2001).
17. Gunka, K. & Commichau, F. M. Control of glutamate homeostasis in *Bacillus subtilis*: a complex interplay between ammonium assimilation, glutamate biosynthesis and degradation. *Mol. Microbiol.* **85,** 213–224 (2012).
18. Stannek, L. *et al.* Evidence for synergistic control of glutamate biosynthesis by glutamate dehydrogenases and glutamate in *Bacillus subtilis*. *Environ. Microbiol.* http://dx.doi.org/10.1111/1462-2920.12813 (2015).
19. Belitsky, B. R. & Sonenshein, A. L. Role and regulation of *Bacillus subtilis* glutamate dehydrogenase genes. *J. Bacteriol.* **180,** 6298–6305 (1998).
20. Zeigler, D. R. *et al.* The origins of 168, W23, and other *Bacillus subtilis* legacy strains. *J. Bacteriol.* **190,** 6983–6995 (2008).
21. Nakano, M. M., Yang, F., Hardin, P. & Zuber, P. Nitrogen regulation of *nasA* and the *nasB* operon, which encode genes required for nitrate assimilation in *Bacillus subtilis*. *J. Bacteriol.* **177,** 573–579 (1995).
22. Kleiner, D. Bacterial ammonium transport. *FEMS Microbiol. Lett.* **32,** 87–100 (1985).
23. Castorph, H. & Kleiner, D. Some properties of a *Klebsiella pneumoniae* ammonium transport negative mutant (Amt-). *Arch. Microbiol.* **139,** 245–247 (1984).
24. Boogerd, F. C. *et al.* AmtB-mediated NH3 transport in prokaryotes must be active and as a consequence regulation of transport by GlnK is mandatory to limit futile cycling of NH4+/NH3. *FEBS Lett.* **585,** 23–28 (2011).
25. Jayakumar, A., Schulman, I., MacNeil, D. & Barnes, E. M. Jr. Role of the *Escherichia coli glnALG* operon in regulation of ammonium transport. *J. Bacteriol.* **166,** 281–284 (1986).
26. Kim, M. *et al.* Need-based activation of ammonium uptake in *Escherichia coli*. *Mol. Syst. Biol.* **8,** 616 (2012).
27. Commichau, F. M., Gunka, K., Landmann, J. J. & Stulke, J. Glutamate metabolism in *Bacillus subtilis*: gene expression and enzyme activities evolved to avoid futile cycles and to allow rapid responses to perturbations of the system. *J. Bacteriol.* **190,** 3557–3564 (2008).
28. Detsch, C. & Stulke, J. Ammonium utilization in *Bacillus subtilis*: transport and regulatory functions of NrgA and NrgB. *Microbiology* **149,** 3289–3297 (2003).
29. Anyan, M. E. *et al.* Type IV pili interactions promote intercellular association and moderate swarming of Pseudomonas aeruginosa. *Proc. Natl Acad. Sci. USA* **111,** 18013–18018 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to thank K. Süel, T. Çağatay, R. Wollman, T. Hwa and M. Elowitz for comments during the writing of the manuscript. A.P. is a Simons Foundation Fellow of the Helen Hay Whitney Foundation. J.H. is supported by the UCSD Cell and Molecular Genetics Training Grant. J.G.-O. is supported by the Ministerio de Economia y Competitividad (Spain) and FEDER, under project FIS2012-37655-C02-01, and by the ICREA Academia Programme. This research was funded by the National Institutes of Health, National Institute of General Medical Sciences Grant R01 GM088428 and the National Science Foundation Grant MCB-1450867 (both to G.M.S.).

Author Contributions G.M.S., J.L., M.A. and J.G.-O. designed the research; J.L. performed the experiments; J.L., J.H. and M.A. performed the data analysis; M.G.-S. and J.G.-O. performed the mathematical modelling; D.D.L., S.L. and M.A. made the bacteria strains; and G.M.S., A.P., J.L., J.H., M.G.-S. and J.G.-O. wrote the manuscript. All authors discussed the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.M.S. (gsuel@ucsd.edu).

METHODS

Strains and plasmids. All experiments were done using *Bacillus subtilis* NCIB 3610. The wild-type strain was a gift from W. Winkler (University of Maryland)[30] and all other strains were derived from it and verified by sequencing (see Supplementary Information).

Growth conditions. The biofilms were grown using MSgg medium[16]. It contains 5 mM potassium phosphate buffer (pH 7.0), 100 mM MOPS buffer (pH 7.0, adjusted using NaOH), 2 mM $MgCl_2$, 700 μM $CaCl_2$, 50 μM $MnCl_2$, 100 μM $FeCl_3$, 1 μM $ZnCl_2$, 2 μM thiamine HCl, 0.5% (v/v) glycerol and 0.5% (w/v) monosodium glutamate. The MSgg medium was made from stock solutions on the day of the experiment, and the stock solution for glutamate was made new each week.

Microfluidics. We used the CellASIC ONIX Microfluidic Platform and the Y04D microfluidic plate (EMD Millipore). It provides unconventionally large chambers, allowing the formation of colonies containing millions of cells, yet still leaves room for media flow. Media flow in the microfluidic chamber was driven by a pneumatic pump from the CellASIC ONIX Microfluidic Platform, and the pressure from the pump was kept stable during the course of the oscillation. In most of the experiments, we used a pump pressure of 1 psi with only one media inlet open (there are 6 media inlets in the Y04D plate), which corresponds to a flow speed of $\sim 16\,\mu\text{m}\,\text{s}^{-1}$ in the growth chamber.

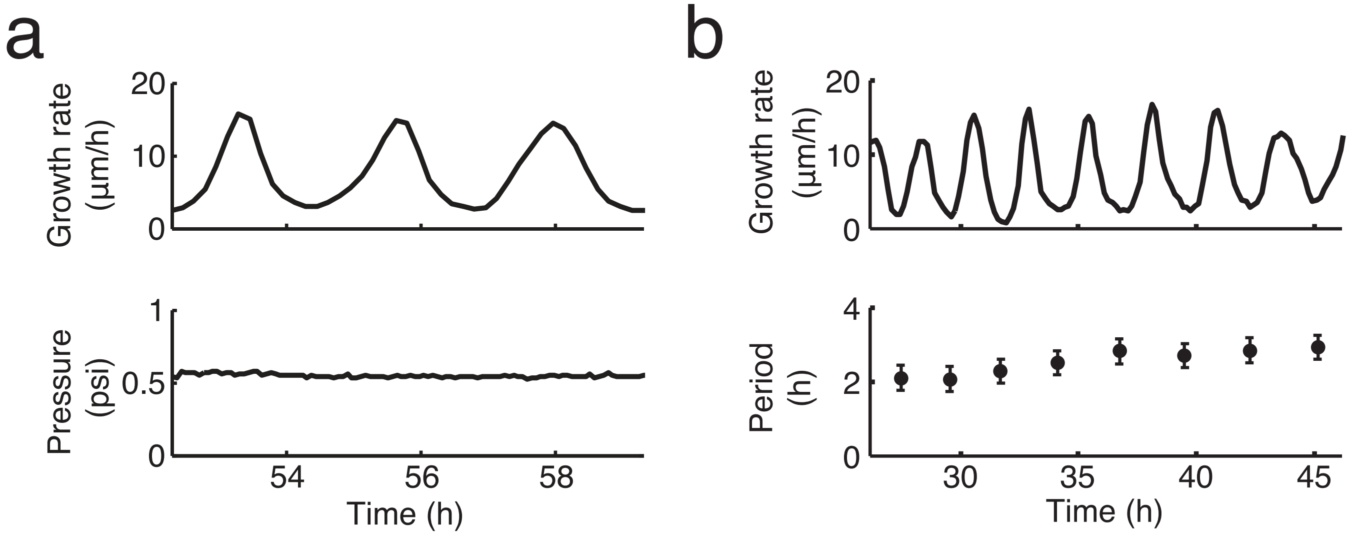
On the day before the experiment, cells from −80 °C glycerol stock were streaked onto LB agar plate and incubated at 37 °C overnight. The next day, a single colony was picked from the plate and inoculated into 3 ml of LB broth in a 50 ml conical tube, and then incubated at 37 °C in a shaker. After 2.5 h of incubation, the cell culture was centrifuged at a relative centrifugal force of 2,100 for 1 min, and then the cell pellet was re-suspended in MSgg and then immediately loaded into microfluidics. After the loading, cells in the microfluidic chamber were incubated at 37 °C for 90 min, and then the temperature was kept at 30 °C for the rest of the experiment.

Time-lapse microscopy. The growth of the biofilms was recorded using phase contrast microscopy. The microscopes used were Olympus IX81 and IX83, and DeltaVision PersonalDV. To image entire biofilms, 10× lens objectives were used in most of the experiments. Images were taken every 10 min. Whenever fluorescence images were recorded, we used the minimum exposure time that still provided a good signal-to-noise ratio.

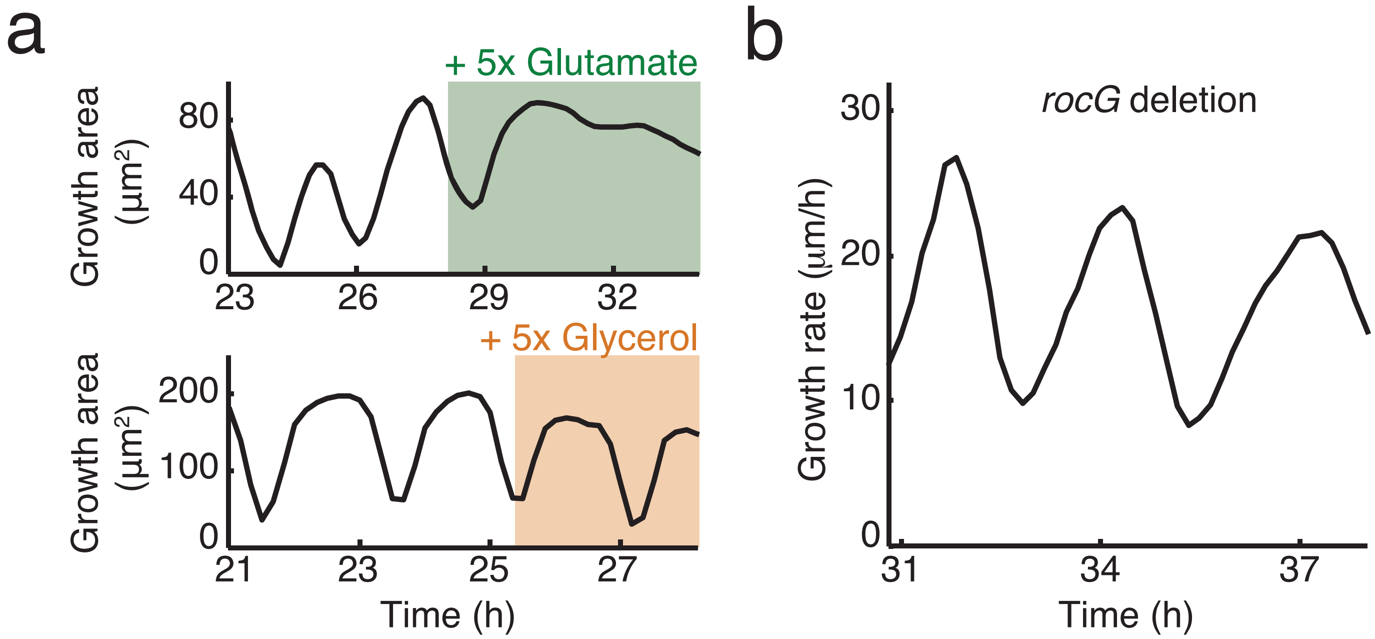
Data analysis. ImageJ (National Institutes of Health) and MATLAB (MathWorks) were used for image analysis. In-house software was also developed to perform colony detection and quantification of colony expansion. Multiple methods of colony detection were used to ensure the accuracy of the analysis. To detect regions of expansion in a biofilm, we performed image differencing on snapshots of the biofilm from time-lapse microscopy videos. Specifically, we calculated the difference between two consecutive phase contrast images (taken 10 min apart) by finding the absolute difference between each pixel in each image. We then generated an image stack based on these results. The intensity values from the stack correlate with the expansion inside the biofilm. The growth area was determined by converting difference images to binary images and then measuring the area of the colony growth region (white pixels). To measure cell replication time, we tracked the length and division of individual cells in the biofilm periphery (Extended Data Fig. 4b).

No statistical methods were used to predetermine sample size.

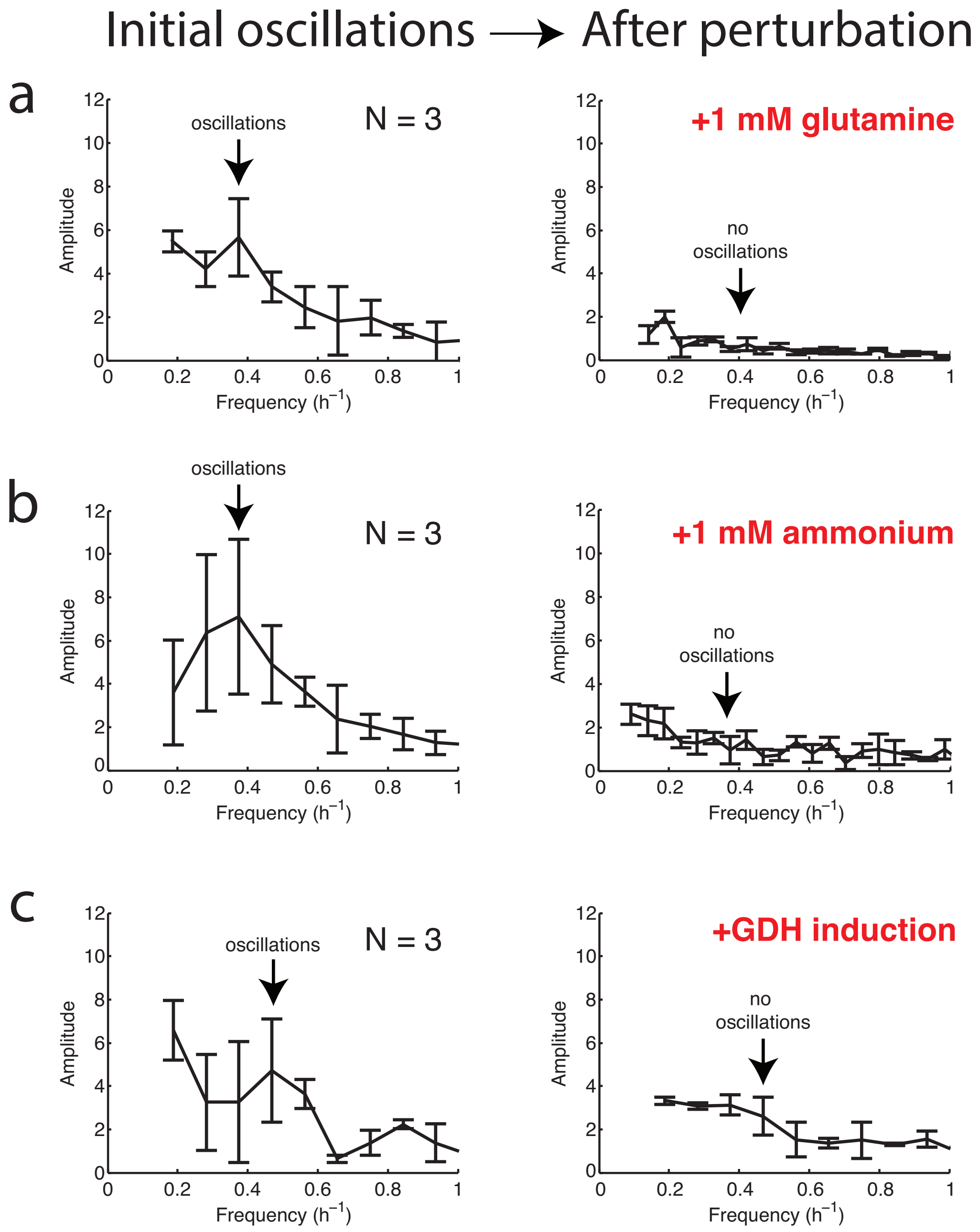
30. Irnov, I. & Winkler, W. C. A regulatory RNA required for antitermination of biofilm and capsular polysaccharide operons in Bacillales. *Mol. Microbiol.* **76,** 559–575 (2010).



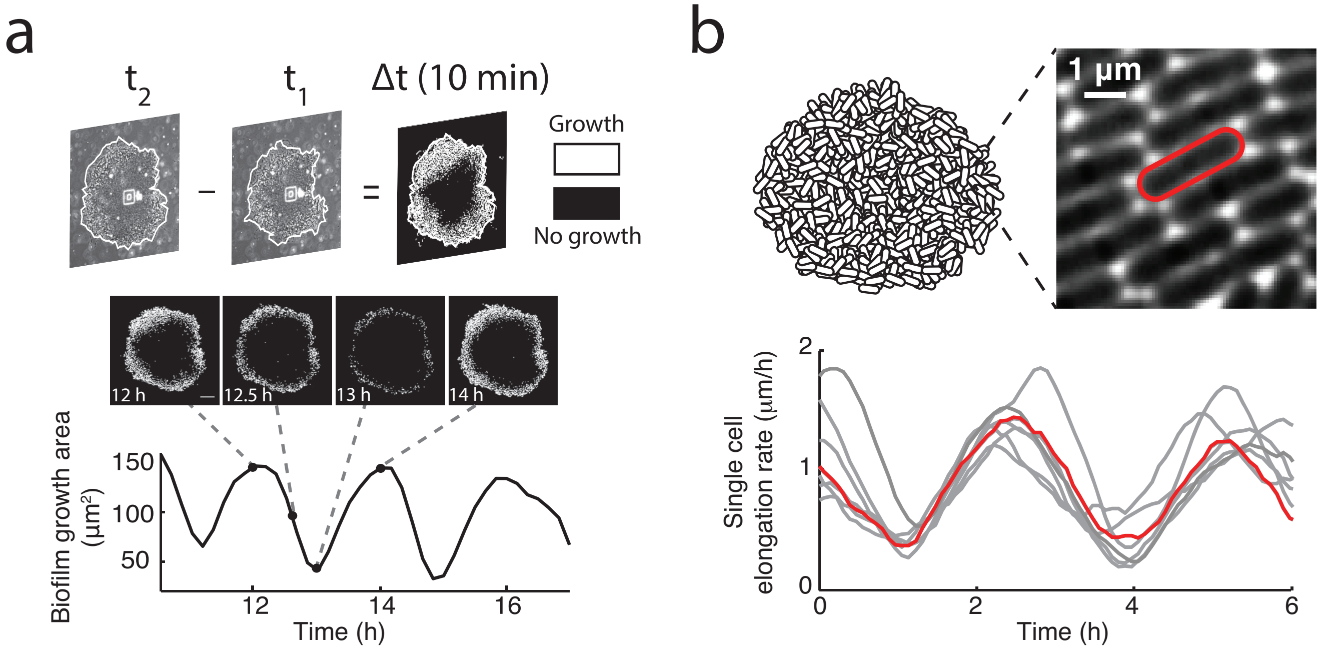
Extended Data Figure 1 | Characterization of biofilm growth oscillations. **a**, Top: growth rate over time of an oscillating colony. Bottom: the pressure that drives media flow in the microfluidic chamber is constant over time (see Methods, 'Microfluidics' section). **b**, Top: growth rate of an oscillating colony. Bottom: period of each oscillation cycle, measured peak to peak. The error bars (±20 min) are determined by the imaging frequency (1 frame per 10 min). The period slightly increases over time (see also Extended Data Fig. 6f and Supplementary Information, 'Mathematical Model').



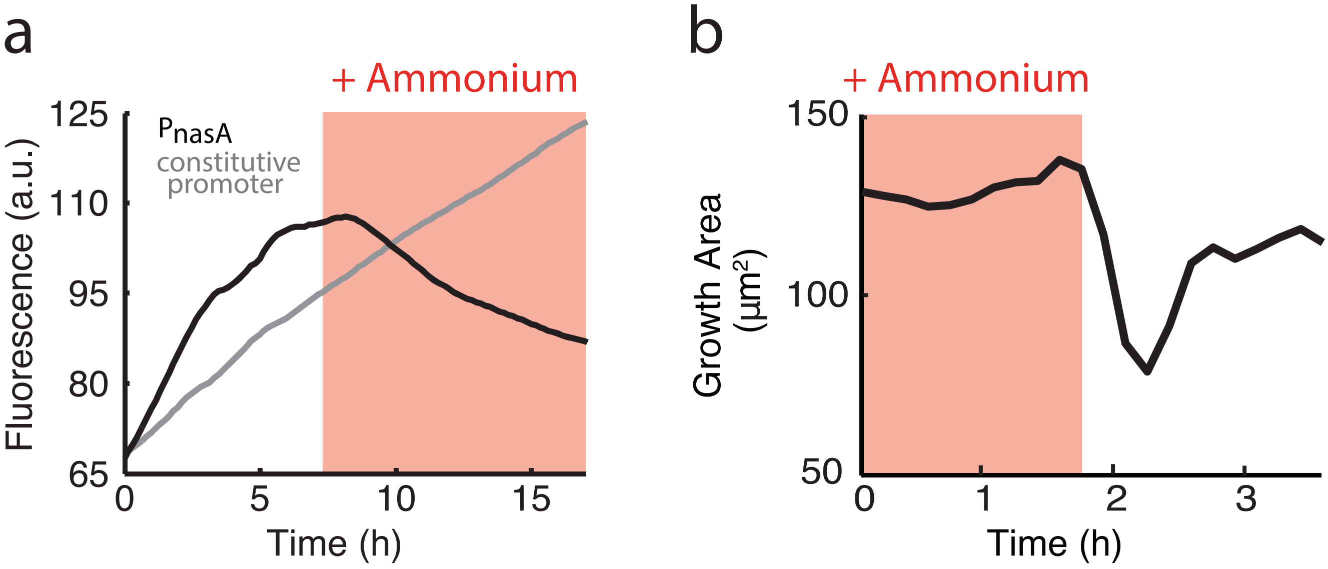
Extended Data Figure 2 | Roles of carbon and nitrogen in biofilm growth oscillations. **a**, Effect of increasing carbon (glycerol) or nitrogen (glutamate) availability on the oscillations. While increasing glutamate by five times of the normal MSgg levels leads to quenching of the oscillation, increasing glycerol by five times does not. **b**, Colony growth of mutant strain with *rocG* deletion. *Bacillus subtilis* NCIB 3610 has two glutamate dehydrogenases (GDH), *rocG* and *gudB*. While *gudB* is constitutively expressed, *rocG* expression is subject to carbon catabolite repression[18]. The oscillatory growth of the *rocG* deletion strain indicates that carbon-source-dependent regulation of *rocG* expression is not required for biofilm oscillations.



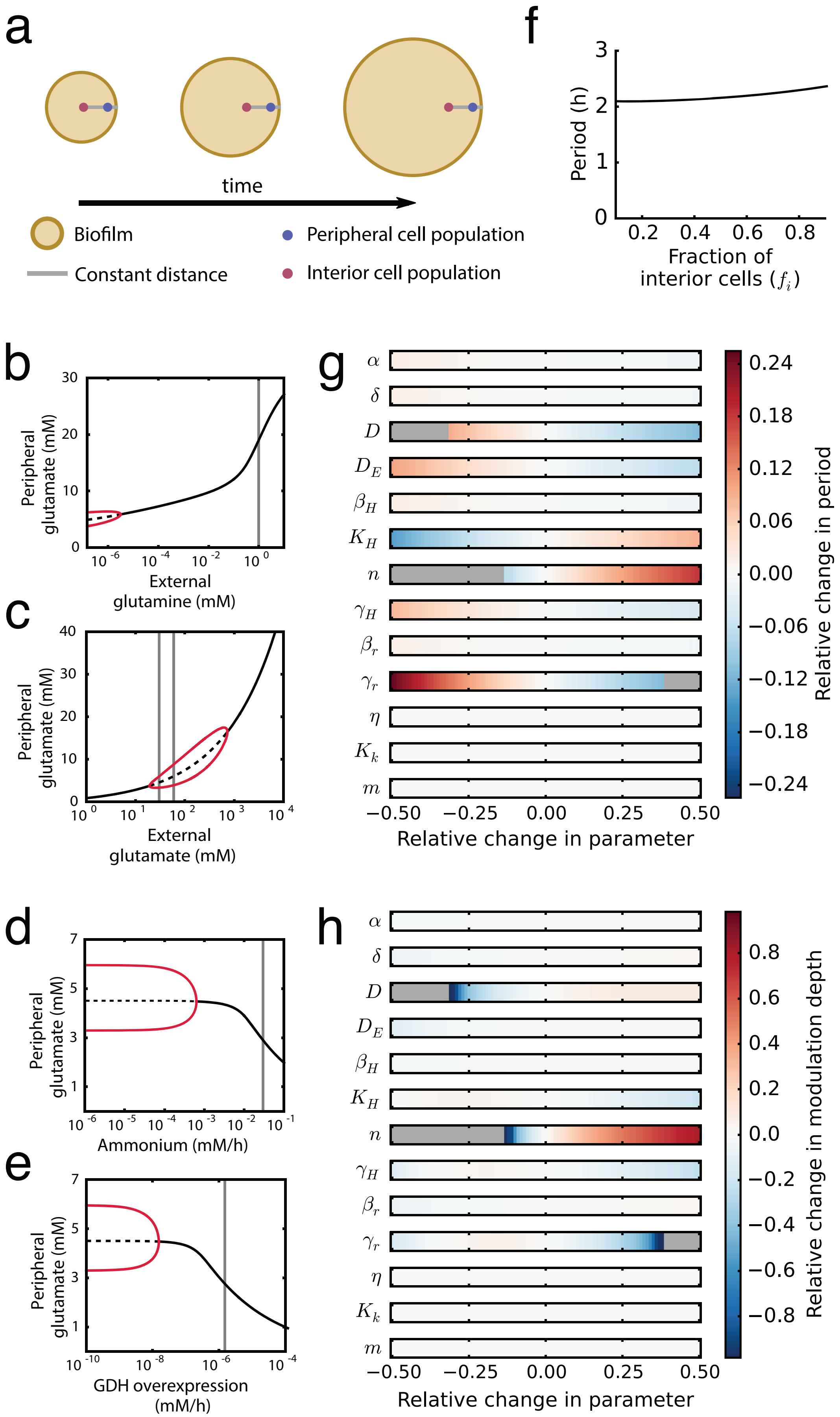
Extended Data Figure 3 | Fourier transform of biofilm growth rates before and after perturbations. The perturbations are: **a**, addition of 1 mM glutamine; **b**, addition of 1 mM ammonium; and **c**, addition of 1 mM IPTG to induce P*hyperspank*-RocG. The error bars show standard deviations (*n* = 3 colonies for each condition). The arrows indicate the frequency of oscillations for each condition before perturbation (left) and the lack of oscillations after perturbation (right).



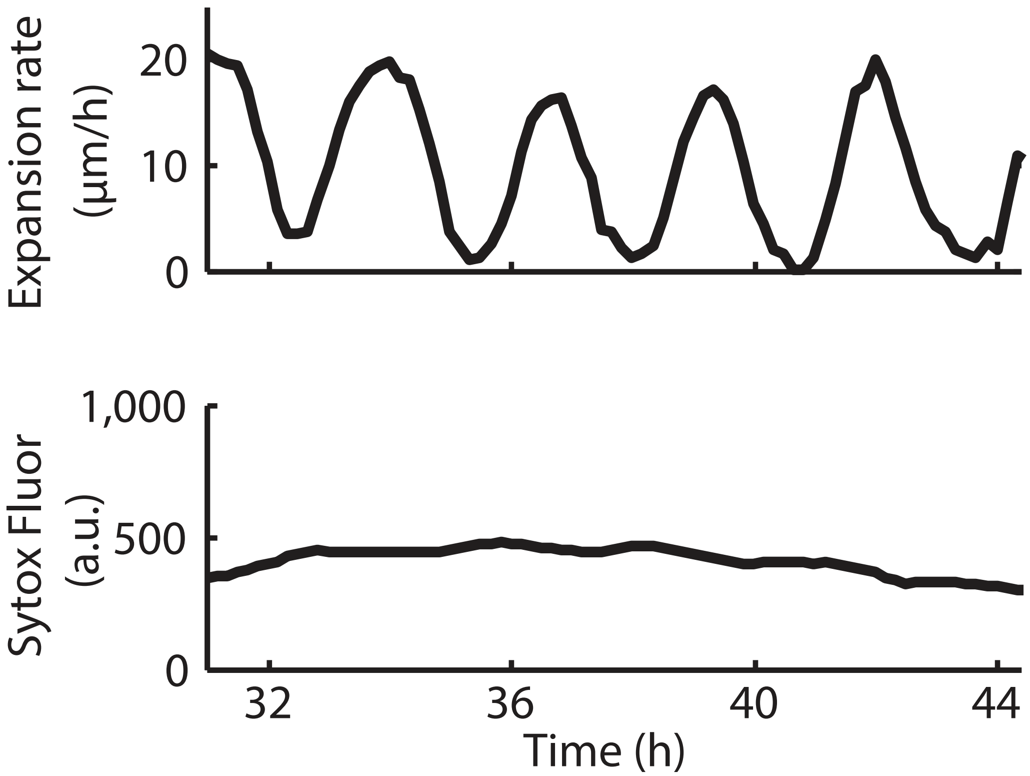
Extended Data Figure 4 | Measurements of cell growth within oscillating biofilms. **a**, Top: visual representation of the method through which difference movies are generated (Methods, 'Data analysis' section). Growth is represented by white pixels, and lack of growth is indicated by black pixels. Film strip (middle) and growth area over time (bottom) of an oscillating colony. Dashed lines show the position of each image on the time trace. Scale bar, 100 μm. **b**, Top left: schematic of a biofilm. Top right: high magnification phase contrast image of biofilm periphery focused at the bottom layer of cells. Bottom panel: time traces depicting elongation rates of single cells in grey. Highlighted in red is the single cell time trace for the cell outlined in red in the top-right panel. The periodic slowdown of the growth of individual peripheral cells is responsible for the observed periodic reduction in biofilm expansion.



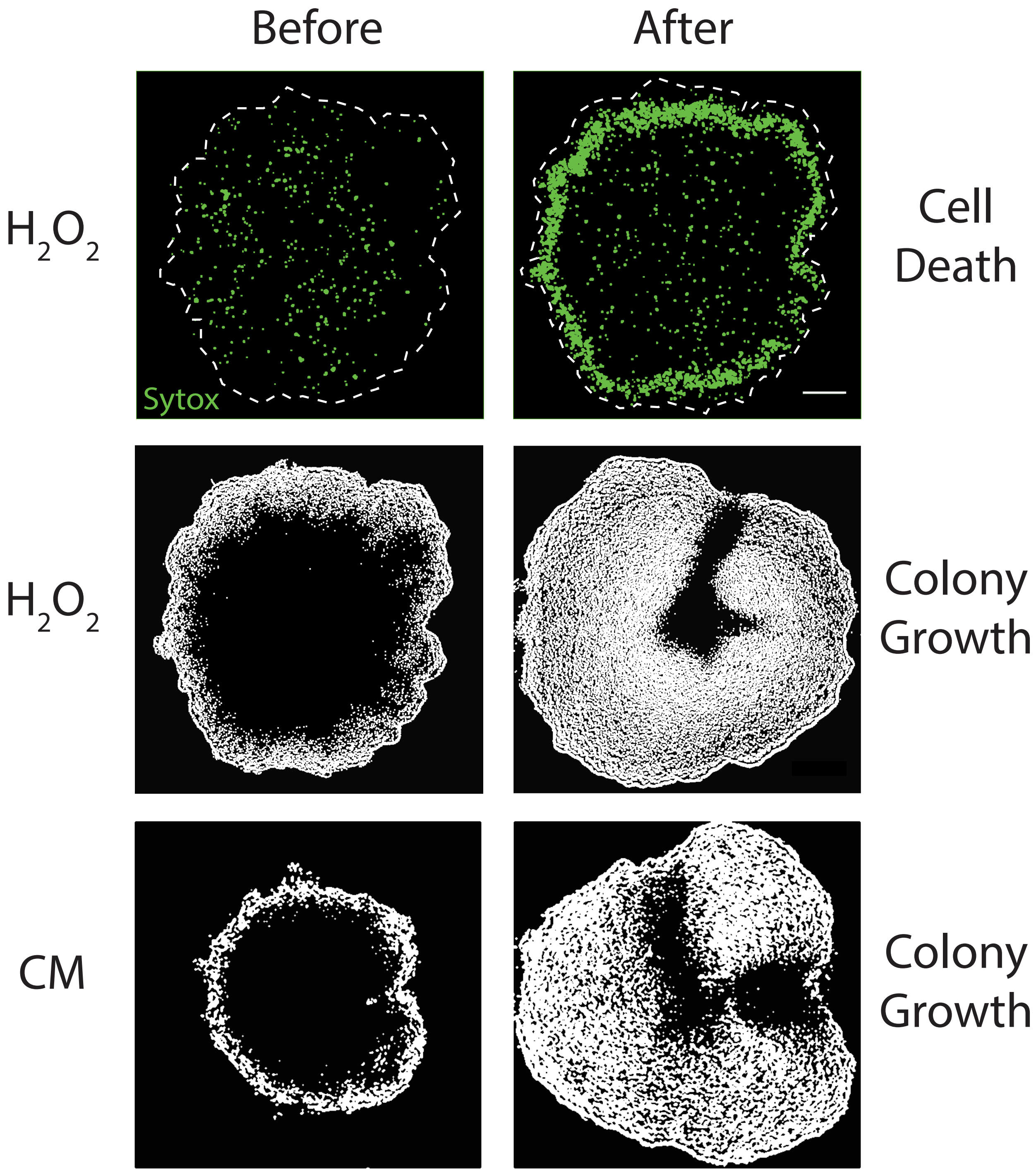
Extended Data Figure 5 | Effects of external ammonium on biofilm development. **a**, Addition of external ammonium (red shading, 1 mM) represses expression from the P*nasA*-YFP reporter (black), but does not affect expression from a constitutive reporter (P*hyperspank*-CFP + 1 mM IPTG, grey). **b**, Removal of external ammonium (red shading, 13 mM) causes halting of colony growth.


a
time
Biofilm
Constant distance
Peripheral cell population
Interior cell population
f
Period (h)
0
1
2
3
0.2
0.4
0.6
0.8
Fraction of interior cells (f_i)
b
Peripheral glutamate (mM)
0
10
20
30
10^{-6}
10^{-4}
10^{-2}
10^{0}
External glutamine (mM)
c
Peripheral glutamate (mM)
0
10
20
30
40
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
External glutamate (mM)
g
α
δ
D
D_E
β_H
K_H
n
γ_H
β_r
γ_r
η
K_k
m
−0.50
−0.25
0.00
0.25
0.50
Relative change in parameter
0.24
0.18
0.12
0.06
0.00
−0.06
−0.12
−0.18
−0.24
Relative change in period
d
Peripheral glutamate (mM)
1
3
5
7
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
Ammonium (mM/h)
e
Peripheral glutamate (mM)
1
3
5
7
10^{-10}
10^{-8}
10^{-6}
10^{-4}
GDH overexpression (mM/h)
h
α
δ
D
D_E
β_H
K_H
n
γ_H
β_r
γ_r
η
K_k
m
−0.50
−0.25
0.00
0.25
0.50
Relative change in parameter
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
Relative change in modulation depth

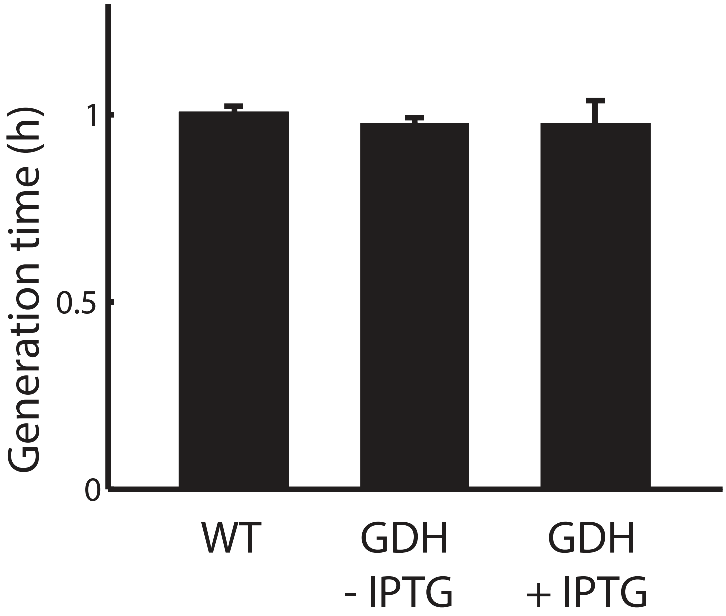
Extended Data Figure 6 | Mathematical model of biofilm growth. a, The model describes the dynamics of two cell populations in a biofilm, interior and peripheral. As the biofilm grows, there is a constant distance between the interior population and the biofilm edge. **b–e**, Bifurcation diagrams showing systematic analysis on the effects of external glutamine, external glutamate, ammonium uptake, and GDH overexpression, respectively. The red lines correspond to the extrema of oscillations in peripheral glutamate (stable limit cycle). The solid black line denotes stable fixed point. The dashed black line corresponds to an unstable fixed point. The vertical grey lines highlight the state of the system for each nutrient addition experiment shown in Fig. 3. **f**, Model prediction of oscillation period as function of interior cell fraction in the whole biofilm. **g**, **h**, Sensitivity analysis of oscillation period (**g**) and modulation depth (**h**) to changes in model parameters. Modulation depth is defined as the amplitude of the oscillations divided by the mean value. Grey colour denotes parameter regions where the system does not oscillate.



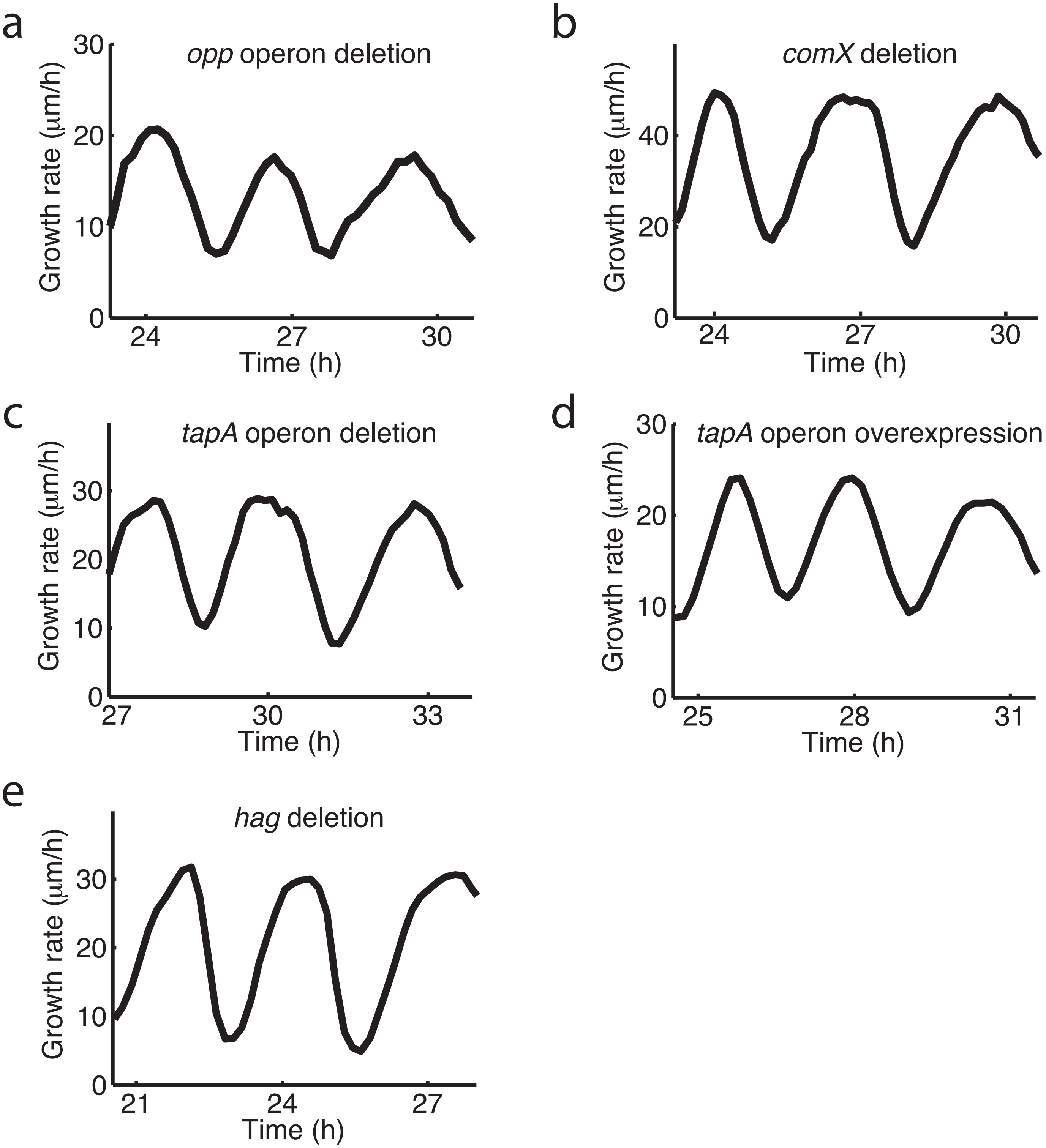
Extended Data Figure 7 | Temporal profile of cell death within an oscillating biofilm. Top: colony growth rate. Bottom: average fluorescence intensity of a cell death marker (Sytox green, 1 µM, Life Technologies) from the same colony shown in the top panel.



Extended Data Figure 8 | Effect of external attack with hydrogen peroxide (H_2O_2, 0.15% v/v) or chloramphenicol (CM, 5 μg ml^{-1}). Top: cell death shown by Sytox green (1 μM). Middle and bottom: colony growth shown by image differencing (see Extended Data Fig. 4a and Methods, 'Data analysis'). Scale bar, 100 μm. The white dashed lines indicate colony edge.



Extended Data Figure 9 | Effect of GDH induction on cell growth. Wild-type and P*hyperspank*-RocG (uninduced or induced with 10 mM IPTG) strains were grown in liquid culture (MSgg medium, 30 °C). Cell generation times were measured using OD_{600}. Error bars show standard deviations ($n = 3$ replicates).



Extended Data Figure 10 | Growth rate oscillations persist in various mutant strains. a, *opp* operon deletion (deficient in quorum sensing). **b**, *comX* deletion (deficient in quorum sensing). **c**, *tapA* operon deletion (extracellular matrix component deletion). **d**, *tapA* operon overexpression (P*hyperspank*-tapA operon, 1 mM IPTG). **e**, *hag* deletion (deficient in swimming and swarming). These results show that the corresponding genes and processes are not required for biofilm oscillations.

doi:10.1038/nature14667

Biogenesis and structure of a type VI secretion membrane core complex

Eric Durand[1,2,3,4,5]*, Van Son Nguyen[2,5]*, Abdelrahim Zoued[1]*, Laureen Logger[1], Gérard Péhau-Arnaudet[4], Marie-Stéphanie Aschtgen[1], Silvia Spinelli[2,5], Aline Desmyter[2,5], Benjamin Bardiaux[4,6], Annick Dujeancourt[3,4], Alain Roussel[2,5], Christian Cambillau[2,5], Eric Cascales[1] & Rémi Fronzes[3,4]

Bacteria share their ecological niches with other microbes. The bacterial type VI secretion system is one of the key players in microbial competition, as well as being an important virulence determinant during bacterial infections. It assembles a nano-crossbow-like structure in the cytoplasm of the attacker cell that propels an arrow made of a haemolysin co-regulated protein (Hcp) tube and a valine–glycine repeat protein G (VgrG) spike and punctures the prey's cell wall. The nano-crossbow is stably anchored to the cell envelope of the attacker by a membrane core complex. Here we show that this complex is assembled by the sequential addition of three type VI subunits (Tss)—TssJ, TssM and TssL—and present a structure of the fully assembled complex at 11.6 Å resolution, determined by negative-stain electron microscopy. With overall C_5 symmetry, this 1.7-megadalton complex comprises a large base in the cytoplasm. It extends in the periplasm via ten arches to form a double-ring structure containing the carboxy-terminal domain of TssM ($TssM_{ct}$) and TssJ that is anchored in the outer membrane. The crystal structure of the $TssM_{ct}$–TssJ complex coupled to whole-cell accessibility studies suggest that large conformational changes induce transient pore formation in the outer membrane, allowing passage of the attacking Hcp tube/VgrG spike.

In the environment, bacteria have evolved collaborative or aggressive mechanisms to communicate, exchange information and chemicals, or compete for space and resources[1–3]. One of the main weapons of bacterial conflicts is a multi-protein device called the type VI secretion system (T6SS) that is assembled in the attacker bacterium[4]. The T6SS is a versatile nanomachine that can deliver toxin proteins directly not only into prey prokaryotes but also into eukaryotic cells during bacterial infections[3,5–9]. Anti-host activities have been shown to inhibit phagocytosis and therefore to disable macrophages, while the anti-bacterial activities allow the bacterium to destroy competitors and to have a privileged access to the niche, to nutrients or to new DNA[3,9,10]. The T6SS is composed of 13 different proteins, encoded by genes that are usually clustered[11]. It assembles a tubular puncturing device that is evolutionarily, structurally and functionally similar to the tail of contractile bacteriophages. Its sheath is a tubular structure, hundreds of nanometres long, that extends in the cytoplasm and is built by the polymerization of TssBC building blocks[12–14]. It is assembled on an assembly platform, the baseplate[13,15–17], and maintained in an extended, metastable conformation[16–18]. The attacking arrow, wrapped by the sheath, comprises an inner tube that is built by stacked Hcp hexameric rings[19] and tipped by a puncturing spike composed of VgrG[20]. Upon contact with the prey, structural rearrangements of the sheath subunits induce its contraction and propulsion of the Hcp tube/VgrG spike towards the target cell, allowing toxin delivery[16,17,21]. The phage-like T6SS tail is anchored to the attacker cell membrane by a trans-envelope complex[22]. This membrane complex not only serves as a docking station but has been proposed as a channel for the passage of the inner tube after sheath contraction, thereby preventing membrane damage in the attacker[16,17]. The membrane core complex of the T6SS (that is, the minimal module required to function and conserved in all T6SS) is composed of the TssL and TssM inner-membrane proteins and the TssJ outer membrane lipoprotein[15–17,22–26]. These proteins are connected through a network of interactions between TssM and TssL, and TssM and TssJ[22,24,25]. Although the localization and topology of these subunits, their interactions and the crystal structures of the soluble domains of TssJ and TssL have been described[17,22–29], we still lack crucial information on the biogenesis and overall architecture of this complex and how it is used as a channel during T6SS action.

Localization, dynamics and biogenesis of the T6SS membrane core complex

We first sought to determine the assembly pathway of the entero-aggregative *Escherichia coli* (EAEC) T6SS membrane core complex. Strains producing fluorescently labelled T6SS membrane subunits were engineered. The sequence encoding the super-folder green fluorescent protein (sfGFP) was inserted upstream of the stop codon of the *tssJ* gene or downstream of the start codon of the *tssL* and *tssM* genes. In these constructs, the fusion proteins were produced from their native chromosomal loci. Hcp release and anti-bacterial assays demonstrated that the sfGFP–TssL and sfGFP–TssM fusion proteins were functional (Extended Data Fig. 1a). By contrast, strains producing TssJ–sfGFP or TssJ–mCherry had a non-functional T6SS (Extended Data Fig. 1b). Fluorescence microscopy analyses showed that sfGFP–TssL and sfGFP–TssM cluster at discrete positions at the cell periphery, in agreement with their membrane localization (Fig. 1a and Extended Data Fig. 1c). These foci are stable and static (Fig. 1a and Extended Data Fig. 1d). Statistical analyses further showed that one or two foci are observable in cells expressing the T6SS (Fig. 1b) and that these clusters are randomly distributed around the cell (Fig. 1c). Co-localization experiments with strains producing

[1] Laboratoire d'Ingénierie des Systèmes Macromoléculaires, Aix-Marseille Université - CNRS, UMR 7255, 31 Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France. [2] Architecture et Fonction des Macromolécules Biologiques, CNRS, UMR 7257, Campus de Luminy, Case 932, 13288 Marseille Cedex 09, France. [3] G5 Biologie structurale de la sécrétion bactérienne, Institut Pasteur, 25–28 rue du Docteur Roux, 75015 Paris, France. [4] UMR 3528, CNRS, Institut Pasteur, 25–28 rue du Docteur Roux, 75015 Paris, France. [5] AFMB, Aix-Marseille Université, IHU Méditerranée Infection, Campus de Luminy, Case 932, 13288 Marseille Cedex 09, France. [6] Unité de Bioinformatique Structurale, Institut Pasteur, 25–28 rue du Docteur Roux, 75015 Paris, France.
*These authors contributed equally to this work.

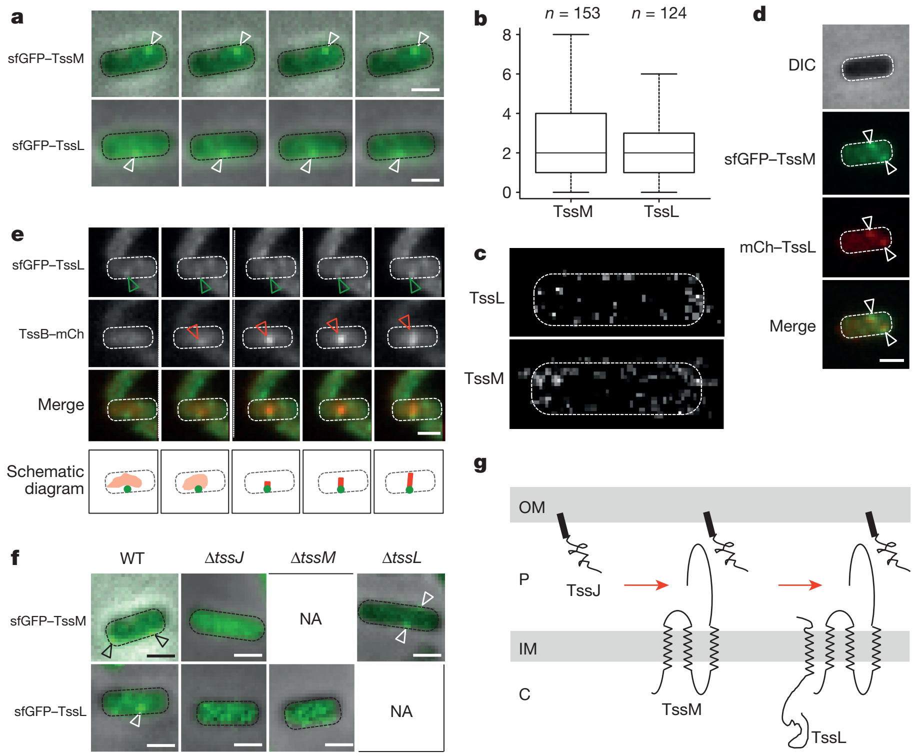


Figure 1 | Biogenesis of the T6SS membrane-associated core complex. **a**, Time-lapse fluorescence microscopy recordings showing localization and dynamics of the sfGFP–TssM and sfGFP–TssL fusion proteins. Individual images were taken every 30 s. The positions of the foci are indicated by arrowheads. Scale bars, 1 µm. Larger fields are presented in Extended Data Fig. 1c. **b**, Statistical analysis of sfGFP–TssM and sfGFP–TssL localization. Shown are box-and-whisker plots of the measured number of sfGFP–TssM and sfGFP–TssL foci per cell for each strain with the lower and upper boundaries of the boxes corresponding to the 25% and 75% percentiles respectively. The black horizontal bar represents the median values for each strain and the whiskers represent the 10% and 90% percentiles. The number of cells studied per strain is indicated above the bars. **c**, Spatial repartition of the sfGFP–TssM and sfGFP–TssL foci. Shown is a superposition of the different foci analysed in a single cell. **d**, sfGFP–TssM and mCh–TssL proteins co-localize. Fluorescence microscopy recordings showing co-localization between sfGFP–TssM and mCh–TssL fusion proteins. The positions of the foci are indicated by the arrowheads. Scale bar, 1 µm. **e**, The membrane complex serves as a docking site for tail sheath polymerization. Time-lapse fluorescence microscopy recordings showing co-localization between sfGFP–TssL and TssB–mCh fusion proteins. Individual images were taken every 30 s. Assembly/contraction of the sheath and TssL localization events is schematized in the bottom row of panels. Scale bars, 1 µm. **f**, Assembly pathway of the T6SS TssJLM membrane complex. Fluorescence microscopy recordings showing sfGFP–TssM and sfGFP–TssL localization in the absence of the TssJ or TssL and TssJ or TssM proteins respectively. The positions of the foci are indicated by the arrowheads. Scale bars, 1 µm. The quantification of the sfGFP–TssM and sfGFP–TssL clusters per cell is presented in Extended Data Fig. 1g. **g**, Schematic representation of the sequential biogenesis of the T6SS membrane complex. The names of the proteins, their localizations and topologies are shown.

sfGFP–TssM and mCherry-tagged TssL showed that the two subunits are present in the foci, demonstrating that each focus corresponds to the position of an assembled membrane complex (Fig. 1d). To test whether these foci are used to anchor the phage-like tail tubular structure, mCherry was fused to the *tssB* sheath gene, at its original chromosomal locus in the strain producing sfGFP–TssL. Time-lapse recordings showed that T6SS sheathes polymerize and extend from the membrane complex (Fig. 1e). On the basis of these results, we conclude that membrane complexes comprising TssL and TssM (and probably TssJ) assemble at discrete positions in the cell and are then used to recruit the tail-complex subunits. Statistical analyses showed that the number of sfGFP–TssL or sfGFP–TssM foci per cell is higher than the number of sheathes (Extended Data Fig. 1e), suggesting that the membrane complexes exist in a pre-assembled form. Interestingly, long-term time-lapse recordings showed that these membrane complexes can be re-used for new tail polymerization events (Extended Data Fig. 1f). To gain further information on the biogenesis of this initial step, *tssJ* or *tssM* were deleted in the sfGFP–TssL-producing strain, and *tssJ* or *tssL* were deleted in the sfGFP–TssM-producing strain. The rationale behind these experiments is that if a protein assembled early is missing, the recruitment of late proteins will be affected, yielding a diffuse fluorescent signal. Figure 1f shows that the recruitment of sfGFP–TssM and sfGFP–TssL is affected in the absence of TssJ, and that of sfGFP–TssL is affected in the absence of TssM. Conversely, the absence of TssL had no effect on TssM recruitment (Fig. 1f and Extended Data Fig. 1g). On the basis of these results, we conclude that TssJ is used as a nucleation factor and that the biogenesis of the T6SS membrane core complex is pursued by the inward sequential addition of TssM and TssL (Fig. 1g).

Architecture of the T6SS membrane core complex

To gain further insights on the architecture of the T6SS membrane core complex, the *tssJ*, *tssL* and *tssM* genes were co-expressed in *E. coli* BL21(DE3). Constructs were designed to add StrepII, Flag and 6×His tags at the carboxy (C) terminus of TssJ, amino (N) terminus of TssL and N terminus of TssM, respectively (Extended Data Fig. 2). Total membranes were isolated and solubilized using detergents. Two-step affinity chromatography followed by gel filtration resulted in the purification of a complex containing TssJ, TssL and TssM (Fig. 2a and Extended Data Fig. 2f–h). In this complex, we determined the TssM–TssL stoichiometry as 1 to 1 (Extended Data Fig. 2h). Purified complexes were visualized by negative-stain electron microscopy (EM) (Fig. 2b and Extended Data Fig. 3a). A data set was collected, and reference-free classification and averaging revealed characteristic views of the complex (class averages) (Fig. 2b). We observed rocket-shaped and ring-shaped views corresponding to side

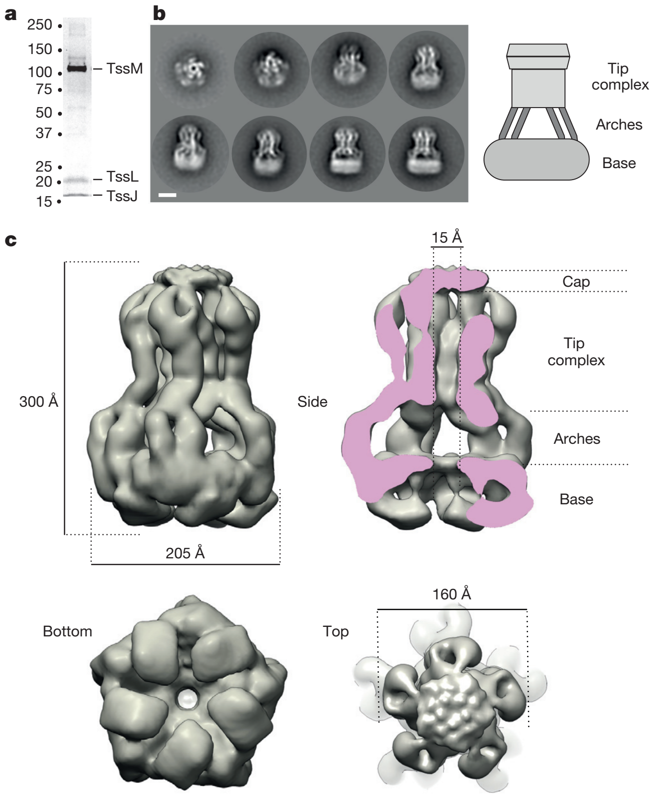


Figure 2 | TssJLM complex purification and structure. a, SDS–polyacrylamide gel electrophoresis (SDS–PAGE) analysis of the purified EAEC TssJLM complex. The bands corresponding to TssM (130 kDa), TssL (24 kDa) and TssJ (18 kDa) after SDS–PAGE and Coomassie blue staining are indicated. **b**, Representative views (class averages) of purified TssJLM complexes. End to side views are shown from top left to bottom right. Scale bar, 10 nm. **c**, Structure of the TssJLM complex. Side, cut-away, bottom and top views are shown from top left to bottom right respectively. The different regions of the complex are indicated on the cut-away view.

and end views of the T6SS membrane core complex respectively (Fig. 2b). Rotational symmetry analysis of end-view class-averages revealed a clear five-fold symmetry in the whole TssJLM population (Extended Data Fig. 3b). The complex is composed of a base and a tip complex linked by arches (Fig. 2b). The negative-stain data set was used to reconstruct a 11.6-Å resolution three-dimensional (3D) volume of the complex with five-fold symmetry applied (Fig. 2c and Extended Data Fig. 3c, d). Local resolution calculations using ResMap[30] indicated that the local resolution was significantly lower in the base (Extended Data Fig. 3e, f). This impaired a correct interpretation of this part of the TssJLM map. Since this could be due to flexibility between the base and the rest of the complex, we performed a local 3D refinement on the base region only, which yielded a 3D reconstruction of the base at 16.6-Å resolution. A composite map where this new reconstruction of the base replaces the equivalent densities in the reconstruction of the whole complex is shown in Fig. 2c. The T6SS membrane core complex is 300 Å in height and 205 Å in diameter (Fig. 2c). It is made of a base that is decorated at its bottom by five hooks and is pierced at its centre by a small hole of 15 Å in diameter (Fig. 2c). Ten arches connect this base to a tip complex of 160 Å in diameter covered by a small cap. Remarkably, five arches gather at the centre of the tip complex to define a 15- to 20-Å diameter channel. The five other arches form a scaffold at the periphery of this complex (Extended Data Fig. 4a). Overall, the tip complex is made of internal and external pillars arranged in concentric rings (Extended Data Fig. 4a).

To define how the core complex is inserted in the cell envelope, we first performed differential solubilization of the inner and outer membranes. The total membrane fraction was solubilized with *N*-lauryl sarkosyl, a detergent that preferentially solubilizes inner-membrane proteins. This differential solubilization resulted in the fractionation of the core complex in both inner and outer membrane fractions (Extended Data Fig. 4b), indicating that this complex resides in both membranes. To determine its orientation in the cell envelope, the purified core complex was incubated with anti-StrepII antibodies or Ni-NTA-coated gold particles targeting the TssJ C-terminal StrepII and TssM N-terminal 6×His tags respectively (Extended Data Fig. 4c), before EM analyses. Anti-StrepII antibodies labelled the tip complex/cap while the base was labelled by the Ni-NTA gold particles (Extended Data Fig. 4c). We concluded that the TssJ C terminus is located in the tip complex while the TssM N terminus is located in the base (Extended Data Fig. 4c). When the N-terminal cysteine residue of the TssJ lipoprotein was substituted by Ser (C1S) to prevent its acylation, an intact TssJ^{C1S}–L–M core complex was formed (Extended Data Fig. 4b), but differential solubilization proved the complex mis-localized to the inner membrane fraction only (Extended Data Fig. 4b). Hence, TssJ acylation tethers the apex of the complex to the outer membrane whereas the base of the complex is located in the cytoplasm.

We next analysed the EM reconstruction to assign the different regions of the core complex to its components. The volume corresponding to one arch and the corresponding pillar within the tip complex (Extended Data Fig. 4a) is comparable in size and shape to that of the isolated TssM periplasmic domain (amino acids 386–1129; TssMp) in complex with TssJ obtained by small-angle X-ray scattering (SAXS) (Extended Data Fig. 4d, e). Segmentation of this volume yielded five different sub-volumes (Fig. 3a). We propose that the sub-volume closest to the cap corresponds to TssJ, in agreement with its location close to the outer membrane. The other four sub-volumes would correspond to sub-domains of TssMp. Sub-volume 4 is in close contact with TssJ, suggesting that it corresponds to the C-terminal domain of TssM domain, which was previously shown to mediate contact with TssJ[25]. With sub-volume 3, it forms the tip complex while sub-volumes 1 and 2 correspond to the arches (Fig. 3a). Interestingly, the last TssM transmembrane segment crossing the inner membrane is located just upstream of TssMp. This would place the inner membrane at the bottom of the arches or at the top of the base. The volume of the base (1,450 Å^3) is much bigger than the estimated volume occupied by ten copies of the cytoplasmic domains of TssM and TssL (825 Å^3). The crystal structure of the TssL cytoplasmic domain dimer[28,29] could be fitted in the hooks with 88% correlation (Extended Data Fig. 4f). This indicates that the remainder of the base could correspond to the cytoplasmic domain of TssM and the 40 transmembrane segments bound to detergent (Extended Data Fig. 4f).

To gain more insight into the structure of the TssMp–TssJ complex, TssMp was produced and purified as described previously[25]. To help crystallization, TssMp complex was subjected to controlled proteolytic digestion[31]. A protease-resistant fragment of an apparent size of ~32 kDa (called hereafter TssM$_{32Ct}$; residues 836–1129; Extended Data Fig. 5a) was further purified and co-crystallized with nb25, a specific camelid single-chain nanobody[31,32]. The structure of the TssM$_{32Ct}$–nb25 complex was solved by molecular replacement using the X-ray structure of nb25 reported previously[32] (Extended Data Fig. 5b and Extended Data Table 1). In the complex, the TssM$_{32Ct}$ amino-acid chains are defined in the electron density map between residues 868 and 1107. We therefore purified a new TssMp fragment (TssM$_{26Ct}$) encompassing the crystallographic visible chain. This shorter fragment crystallized readily alone as well as in complex with the unacylated TssJ protein (Extended Data Table 1). The structure of TssM$_{26Ct}$ is composed of two domains. The N-terminal domain (residues 870–974) is a bundle of four α-helices, covered on one side by a β-hairpin (Fig. 3b) and on the other by the C-terminal elongated stretch of the protein. The C-terminal domain (residues 975–1085) is a nine-stranded β-sandwich

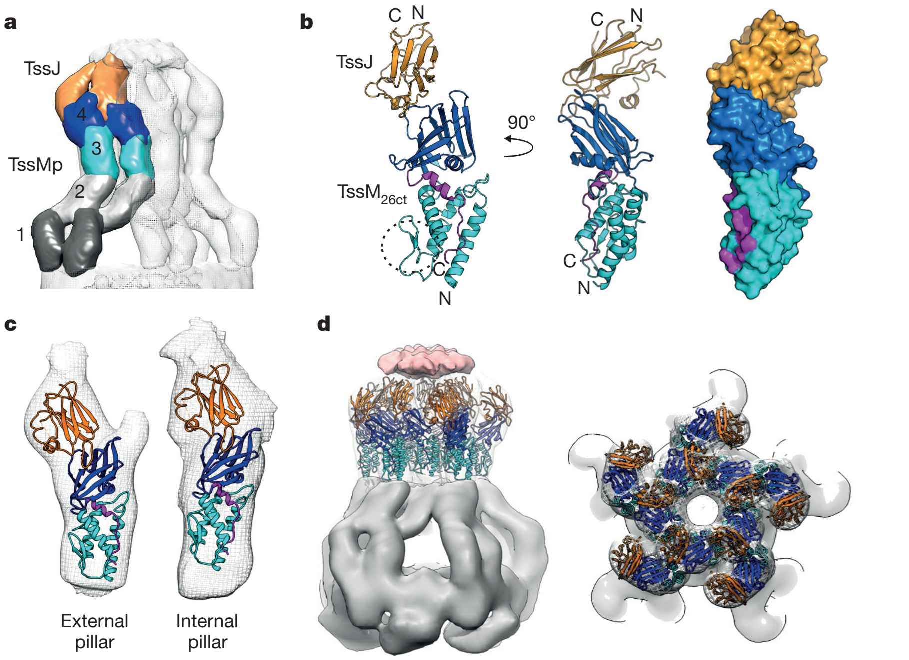


Figure 3 | Structure of the TssJLM tip complex. **a**, Segmentation of the TssJLM complex reconstruction. Each volume encompassing one arch and the corresponding pillar within the tip complex is segmented in five different domains (shown in different colours). **b**, Crystal structure of the $TssM_{26Ct}$–$TssJ_{sol}$ complex represented as ribbons. $TssJ_{sol}$ is coloured orange, while $TssM_{26Ct}$ is coloured cyan (α-domain) and blue (β-domain). The C-terminal α5-helix and the extended stretch are coloured magenta. The β-hairpin (β1–β2) is highlighted in the dashed circle. Two orthogonal views of the crystal structure and its surface representation are shown from left to right (coloured as in **a**). **c**, $TssM_{26Ct}$–TssJ crystal structure docked into the EM volume corresponding to TssJ and the TssM periplasmic domains 3 and 4 extracted from both internal and external pillars of the tip complex. **d**, Energy-minimized atomic model of the tip complex structure (left panel, side view; right panel, top view).

that contacts nb25 or TssJ (Fig. 3b and Extended Table 2a, b). This C-terminal domain is followed by a stretch of residues (1086–1107) comprising helix α5 (Fig. 3b). TssJ binds to the apex of the C-terminal domain, and the 590-$Å^2$ interaction area involves contacts between TssJ loops L1–2, L3–4 and L5–6 with $TssM_{26Ct}$ loops L3–4 and L5–6 (Extended Data Fig. 5c and Extended Data Table 2b), in agreement with a previous study demonstrating the importance of TssJ loop L1–2 for TssM–TssJ complex formation[25]. Superimposition of the structures of $TssM_{32Ct}$–nb25 and $TssM_{26Ct}$–TssJ shows that nb25 and TssJ cannot bind simultaneously to TssM (Extended Data Fig. 5d), explaining the nb25 *in vivo* inhibitory effect on T6SS function[32]. The comparison between $TssM_{26Ct}$–TssJ crystal structure and the volume proposed to correspond to TssJ and domains 3 and 4 of TssMp determined by EM resulted in 95% correlation between the two structures (Fig. 3c). This confirms the location of $TssM_{26Ct}$–TssJ in the tip complex (Fig. 3d).

Cell surface accessibility and transient pore formation

The orientation of the TssJ N terminus places the outer membrane above TssJ, where the cap is located (Figs 2c and 4a and Extended Data Fig. 5e). Accordingly, close inspection of the proposed oligomeric structure of the $TssM_{26Ct}$–TssJ complex could not reveal any obvious transmembrane region (Extended Data Fig. 5f). To test this, we engineered functional cysteine derivatives between the β-strands of the C-terminal domain of TssM (Extended Data Fig. 6a). The extracellular accessibility of these residues was assessed by incubating whole cells with an outer membrane-impermeant cysteine-reactive maleimide. We observed that positions 989, 1005, 1035, 1075 and 1109 were labelled whereas positions 972, 1019, 1062 and 1092 were not (Extended Data Fig. 6b and Extended Data Table 2c). With the exception of position 1092, all other positions were labelled when cell lysates were used instead of intact cells (Extended Data Table 2c). The labelled cysteine substitutions are on the tip of TssM facing the outer membrane (Fig. 4a). Interestingly, residues 989 and 1005 are buried at the interface with TssJ (Extended Data Fig. 6c). Therefore, for these residues to be labelled, the TssM–TssJ complex has to dissociate. This result also suggests that the tip of $TssM_{26Ct}$ is exposed to the cell exterior. To test whether TssM stably crosses the outer membrane or accesses the cell exterior temporarily, similar experiments were conducted in a *tssBC*-deleted background. In the absence of the TssB and TssC sheath components, the TssJLM membrane complex is properly assembled but the T6SS is inactive as no sheath assembly or contraction could occur. In *tssBC* cells, only position 1109 was labelled (Extended Data Fig. 6b and Fig. 4a). These results suggest that the TssM α5-helix crosses the outer membrane permanently, exposing the C-terminal extension to the extracellular medium whereas part of $TssM_{26Ct}$ domain is exposed transiently at the cell surface during the T6SS mechanism of action.

Closing remarks and outlook

The data presented here allow an unprecedented understanding of the biogenesis, architecture and role of the T6SS TssJLM membrane core complex. This complex anchors the phage tail-like structure to the cell envelope and is thought to serve as conduit to guide the Hcp tube/VgrG spike upon sheath contraction[15–17]. Using fluorescence microscopy, we demonstrate that the three subunits are recruited in a specific order, starting from the outer membrane TssJ lipoprotein and pursued by the sequential addition of TssM and TssL, a hierarchy in agreement with previously published localization and interaction studies[17,22–27]. Therefore, T6SS biogenesis is initiated by an outer membrane lipoprotein nucleation factor and progresses inwards, like the assembly mechanisms of other bacterial secretion systems[33–39]. Our fluorescence microscopy analyses also showed that the T6SS membrane core complex assembles randomly in the cell envelope, without specific localization. The complex is stable and can be used for several events of sheath assembly/contraction, increasing the amount of toxin effectors delivered to the target cell.

The TssJLM complex has a five-fold symmetry and is composed of ten copies of each component that assemble a 1.7-MDa structure crossing the inner membrane, the periplasm and anchored to the outer membrane via the TssJ N-terminal lipid moiety. Its architecture is unique compared with other trans-envelope bacterial secretion systems (Extended Data Fig. 7a). On the basis of our accessibility experiments, we propose that upon assembly of other T6SS subunits with the membrane core complex, the TssM C-terminal extension (C-terminal extended stretch following helix α5 in the crystal structure and the remaining 22 non-visible amino acids) will change its conformation and will cross the outer membrane. The base of the TssJLM complex defines a small cavity and hole that cannot accommodate the VgrG protein and potential effectors bound to it (Fig. 4b, stages 1 and 2)[9,20]. We propose that the base changes its conformation

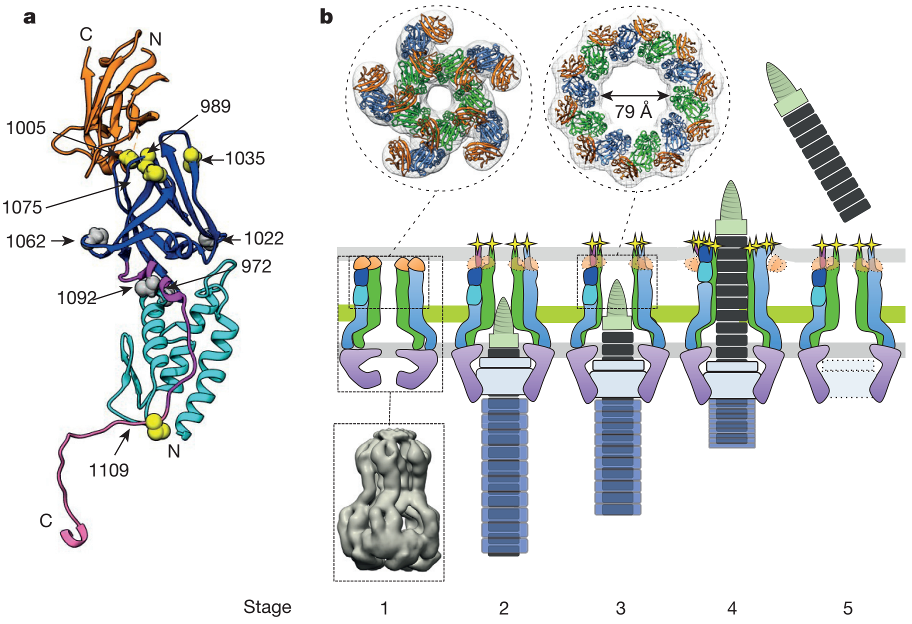


Figure 4 | Cell surface accessibility and mechanism of action of the T6SS membrane core complex during secretion. **a**, Cell surface accessibility studies. Crystal structure of the TssM$_{26Ct}$ represented as ribbons, coloured cyan (α-domain) and blue (β-domain). The C-terminal α5-helix and the extended stretch are coloured magenta. The C terminus (lacking in the crystallized fragment) is represented as a random structure beyond the last residue in the crystallographic model. The cysteine substitutions (in sphere representation) used for labelling experiments are positioned in the TssM$_{26Ct}$ crystal structure. Cysteines with extracellular accessibility when the T6SS is active are coloured yellow, while the unlabelled ones are coloured grey. **b**, Model of action. The proposed mechanism of action involves five sequential stages. Stage 1: the assembled TssJLM complex is not integrally inserted in the outer membrane, but anchored to it by the TssJ N-terminal lipid moiety. This stage corresponds to the EM reconstruction of the purified TssJLM complex (bottom inset) and the crystal structure of the TssM$_{26Ct}$–TssJ complex (top inset) presented in this study. Stage 2: upon assembly of T6SS baseplate and tail components, the C-terminal extremity of TssM inserts into the outer membrane and is therefore accessible at the cell surface (yellow star). This stage corresponds to the 'resting' state of the T6SS membrane complex. Stages 3 and 4: the membrane core complex opens to allow the passage of the Hcp tube/VgrG spike or the sheath contraction force induces conformational changes of the TssJLM complex. A molecular model of a C$_{10}$ symmetrized TssJ/TssM$_{26Ct}$ ring is presented (top inset). The apical loops of TssM are exposed at the cell surface (yellow stars). Stage 5: after release of the Hcp tube/VgrG spike, the TssJLM membrane complex returns to the resting state, ready to perform another cycle of secretion.

to receive the baseplate components. This state would correspond to a 'resting' state of the T6SS machinery (Fig. 4b, stage 2). Ten arches cross the periplasm and are followed by ten pillars positioned in two concentric layers in the tip complex. The inner pillars define a channel of 15–20 Å in diameter that is not large enough to allow the passage of the ~110 Å Hcp tube[16–18] (Extended Data Fig. 7b). Interestingly, it was previously shown that TssM undergoes large conformational changes during secretion[26]. Therefore, we propose that the inner TssM pillars are pushed outwards to define a wider TssM ring with internal dimensions compatible with the passage of the Hcp tube/VgrG spike (Fig. 4b, stages 3 and 4, and Extended Data Fig. 7b, c). In other secretion systems, specific components are dedicated to assemble the outer membrane pore. No obvious transmembrane region could be found in the TssM C-terminal domain or in TssJ. It is unlikely that the C-terminal portion of TssM would form a pore of sufficient dimension by itself. Therefore, we propose that the stroke of the Hcp–VgrG arrow would mechanically push the C-terminal TssM domain towards the cell exterior, allowing the transient formation of a pore through the outer membrane (Fig. 4b, stage 4). To avoid deleterious effects for the bacterium, one may expect that the C-terminal domain of TssM returns to its initial 'resting' conformation at the periplasmic face of the outer membrane once the Hcp tube has been released, closing the outer membrane channel (Fig. 4b, stage 5). Overall, the membrane core complex appears to act like a docking station for the phage-like T6SS device. It nucleates the assembly of the rest of the secretion system and then guides the Hcp tube/VgrG spike through the bacterial cell envelope upon sheath contraction. Further studies will be necessary to fully understand the complete assembly process of the T6SS, the trigger that releases sheath contraction and how the Hcp tube/VgrG spike crosses both bacterial and host membranes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 10 March; accepted 15 June 2015.
Published online 22 July 2015.

1. West, S. A., Griffin, A. S. & Gardner, A. Evolutionary explanations for cooperation. *Curr. Biol.* **17,** 661–672 (2007).
2. Blango, M. G. & Mulvey, M. A. Bacterial landlines: contact-dependent signaling in bacterial populations. *Curr. Opin. Microbiol.* **12,** 177–181 (2009).
3. Russell, A. B., Peterson, S. B. & Mougous, J. D. Type VI secretion system effectors: poisons with a purpose. *Nature Rev. Microbiol.* **12,** 137–148 (2014).
4. Silverman, J. M., Brunet, Y. R., Cascales, E. & Mougous, J. D. Structure and regulation of the type VI secretion system. *Annu. Rev. Microbiol.* **66,** 453–472 (2012).
5. Pukatzki, S., Ma, A. T., Revel, A. T., Sturtevant, D. & Mekalanos, J. J. Type VI secretion system translocates a phage tail spike-like protein into target cells where it cross-links actin. *Proc. Natl Acad. Sci. USA* **104,** 15508–15513 (2007).
6. Russell, A. B. *et al.* Type VI secretion delivers bacteriolytic effectors to target cells. *Nature* **475,** 343–347 (2011).
7. Russell, A. B. *et al.* Diverse type VI secretion phospholipases are functionally plastic antibacterial effectors. *Nature* **496,** 508–512 (2013).
8. Ma, L. S., Hachani, A., Lin, J. S., Filloux, A. & Lai, E. M. *Agrobacterium tumefaciens* deploys a superfamily of type VI secretion DNase effectors as weapons for interbacterial competition in planta. *Cell Host Microbe* **16,** 94–104 (2014).
9. Durand, E., Cambillau, C., Cascales, E., Journet, L. & Vgr, G. Tae, Tle, and beyond: the versatile arsenal of Type VI secretion effectors. *Trends Microbiol.* **22,** 498–507 (2014).
10. Borgeaud, S., Metzger, L. C., Scrignari, T. & Blokesch, M. Bacterial evolution. The type VI secretion system of *Vibrio cholerae* fosters horizontal gene transfer. *Science* **347,** 63–67 (2015).
11. Cascales, E. The type VI secretion toolkit. *EMBO Rep.* **9,** 735–741 (2008).
12. Bönemann, G., Pietrosiuk, A., Diemand, A., Zentgraf, H. & Mogk, A. Remodelling of VipA/VipB tubules by ClpV-mediated threading is crucial for type VI protein secretion. *EMBO J.* **28,** 315–325 (2009).
13. Basler, M., Pilhofer, M., Henderson, G. P., Jensen, G. J. & Mekalanos, J. J. Type VI secretion requires a dynamic contractile phage tail-like structure. *Nature* **483,** 182–186 (2012).

14. Kudryashev, M. *et al.* Structure of the type VI secretion system contractile sheath. *Cell* **160,** 952–962 (2015).
15. Coulthurst, S. J. The type VI secretion system – a widespread and versatile cell targeting system. *Res. Microbiol.* **164,** 640–654 (2013).
16. Ho, B. T., Dong, T. G. & Mekalanos, J. J. A view to a kill: the bacterial type VI secretion system. *Cell Host Microbe* **15,** 9–21 (2014).
17. Zoued, A. *et al.* Architecture and assembly of the type VI secretion system. *Biochim. Biophys. Acta* **1843,** 1664–1673 (2014).
18. Bönemann, G., Pietrosiuk, A. & Mogk, A. Tubules and donuts: a type VI secretion story. *Mol. Microbiol.* **76,** 815–821 (2010).
19. Brunet, Y. R., Hénin, J., Celia, H. & Cascales, E. Type VI secretion and bacteriophage tail tubes share a common assembly pathway. *EMBO Rep.* **15,** 315–321 (2014).
20. Shneider, M. M. *et al.* PAAR-repeat proteins sharpen and diversify the type VI secretion system spike. *Nature* **500,** 350–353 (2013).
21. Brunet, Y. R., Espinosa, L., Harchouni, S., Mignot, T. & Cascales, E. Imaging type VI secretion-mediated bacterial killing. *Cell Rep.* **3,** 36–41 (2013).
22. Aschtgen, M. S., Gavioli, M., Dessen, A., Lloubès, R. & Cascales, E. The SciZ protein anchors the enteroaggregative *Escherichia coli* type VI secretion system to the cell wall. *Mol. Microbiol.* **75,** 886–899 (2010).
23. Aschtgen, M. S., Bernard, C. S., de Bentzmann, S., Lloubès, R. & Cascales, E. SciN is an outer membrane lipoprotein required for type VI secretion in enteroaggregative *Escherichia coli. J. Bacteriol.* **190,** 7523–7531 (2008).
24. Ma, L. S., Lin, J. S. & Lai, E. M. An IcmF family protein, ImpLM, is an integral inner membrane protein interacting with ImpKL, and its walker a motif is required for type VI secretion system-mediated Hcp secretion in *Agrobacterium tumefaciens. J. Bacteriol.* **191,** 4316–4329 (2009).
25. Felisberto-Rodrigues, C. *et al.* Towards a structural comprehension of bacterial type VI secretion systems: characterization of the TssJ-TssM complex of an *Escherichia coli* pathovar. *PLoS Pathog.* **7,** e1002386 (2011).
26. Ma, L. S., Narberhaus, F. & Lai, E. M. IcmF family protein TssM exhibits ATPase activity and energizes type VI secretion. *J. Biol. Chem.* **287,** 15610–15621 (2012).
27. Aschtgen, M. S., Zoued, A., Lloubès, R., Journet, L. & Cascales, E. The C-tail anchored TssL subunit, an essential protein of the enteroaggregative *Escherichia coli* Sci-1 type VI secretion system, is inserted by YidC. *MicrobiologyOpen* **1,** 71–82 (2012).
28. Durand, E. *et al.* Structural characterization and oligomerization of the TssL protein, a component shared by bacterial type VI and type IVb secretion systems. *J. Biol. Chem.* **287,** 14157–14168 (2012).
29. Chang, J. H. & Kim, Y. G. Crystal structure of the bacterial type VI secretion system component TssL from *Vibrio cholerae. J. Microbiol.* **53,** 32–37 (2015).
30. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods* **11,** 63–65 (2014).
31. Nguyen, V. S. *et al.* Production, crystallization and X-ray diffraction analysis of a complex between a fragment of the TssM T6SS protein and a camelid antibody. *Acta Crystallogr F.* **71,** 266–271 (2015).
32. Nguyen, V. S. *et al.* Inhibition of type VI secretion by an anti-TssM llama nanobody. *PLoS ONE* **10,** e0122187 (2015).
33. Diepold, A. *et al.* Deciphering the assembly of the *Yersinia* type III secretion injectisome. *EMBO J.* **29,** 1928–1940 (2010).
34. Judd, P. K., Kumar, R. B. & Das, A. Spatial location and requirements for the assembly of the *Agrobacterium tumefaciens* type IV secretion apparatus. *Proc. Natl Acad. Sci. USA* **102,** 11498–11503 (2005).
35. Hardie, K. R., Lory, S. & Pugsley, A. P. Insertion of an outer membrane protein in *Escherichia coli* requires a chaperone-like protein. *EMBO J.* **15,** 978–988 (1996).
36. Drake, S. L., Sandstedt, S. A. & Koomey, M. PilP, a pilus biogenesis lipoprotein in *Neisseria gonorrhoeae*, affects expression of PilQ as a high-molecular-mass multimer. *Mol. Microbiol.* **23,** 657–668 (1997).
37. Burghout, P. *et al.* Role of the pilot protein YscW in the biogenesis of the YscC secretin in *Yersinia enterocolitica. J. Bacteriol.* **186,** 5366–5375 (2004).
38. Crago, A. M. & Koronakis, V. *Salmonella* InvG forms a ring-like multimer that requires the InvH lipoprotein for outer membrane localization. *Mol. Microbiol.* **30,** 47–56 (1998).
39. Daefler, S. & Russel, M. The *Salmonella typhimurium* InvH protein is an outer membrane lipoprotein required for the proper localization of InvG. *Mol. Microbiol.* **28,** 1367–1380 (1998).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was funded by Agence Nationale de la Recherche (ANR) grants ANR-10-JCJC-1303-03 to E.C., Bip:Bip to R.F., ANR-14-CE14-0006-02 to C.C. and E.C., and supported by the French Infrastructure for Integrated Structural Biology (FRISBI) ANR-10-INSB-05-01. E.D. was supported by a post-doctoral fellowship from the Fondation pour la Recherche Médicale (SPF20101221116) and ANR grants Bip:Bip and ANR-10-JCJC-1303-03. V.S.N. was supported by a PhD grant from the French Embassy in Vietnam (792803C). A.Z., L.L. and M.S.A. were recipients of doctoral fellowships from the French Ministère de la Recherche. A.Z. received a Fondation pour la Recherche Médicale fellowship (FDT20140931060). We thank O. Francetic for providing anti-DglA and anti-OmpF antibodies. We thank E. Marza, P. Violinova Krasteva and H. Remaut for comments on the manuscript, and T. Mignot, M. Guzzo and L. Espinosa for advice about the fluorescence microscopy experiments and the statistical analyses. We also thank the members of the R.F. and E.C. research groups for discussions and suggestions, and R. Lloubès, J. Sturgis and A. Galinier for encouragement. We thank the ERSF and Soleil Synchrotron radiation facilities for beamline allocation.

Author Contributions E.D., A.Z., C.C., E.C. and R.F. designed the experiments. A.Z. constructed the EAEC mutant and fluorescent strains and performed the fluorescence microscopy experiments and statistical analyses. L.L. and M.S.A. constructed the TssM cysteine derivatives and performed the accessibility experiments. E.D. assisted by An.D. purified the TssJLM complex and performed its biochemical characterization. E.D. and G.P.A. collected the EM data. E.D. and R.F. obtained the 3D reconstruction of the TssJLM complex. V.S.N., S.S., A.R. and C.C. purified, crystallized and solved the X-ray structures. Al.D. generated the nanobody. B.B. obtained the energy-minimized models of the closed and open states of the TssM$_{26Ct}$–TssJ complex.

Author Information The EM structure of the TssJLM complex has been deposited in the Electron Microscopy Data Bank (EMDB) under accession number emd-2927. The crystal structures of the TssM$_{32Ct}$–nb25 complex, and of the TssM$_{26Ct}$ fragment and TssM$_{26Ct}$–TssJ complexes, have been deposited in the Protein Data Bank under accession numbers 4Y7M, 4Y7L and 4Y7O, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.C. (cambillau@afmb.univ-mrs.fr), E.C. (cascales@imm.cnrs.fr) or R.F. (remi.fronzes@pasteur.fr).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Strains, media and chemicals. The strains, plasmids and oligonucleotides used in this study are listed in Supplementary Table 1. The *E. coli* K-12 DH5α strain was used for cloning steps whereas *E. coli* K-12 BL21(DE3) and T7-Iq pLys strains were used for protein purification. The enteroaggregative *E. coli* EAEC strain 17-2 was used to engineer gene knockouts and fusions with fluorescent labels. Strains were routinely grown in lysogeny broth (LB) rich medium (or Terrific broth medium for protein purification) or in Sci-1-inducing medium (SIM; M9 minimal medium, glycerol 0.2%, vitamin B1 1 μg ml^{-1}, casaminoacids 100 μg ml^{-1}, LB 10%, supplemented or not with bactoagar 1.5%)[40] with shaking at 37 °C. Plasmids were maintained by the addition of ampicillin (100 μg ml^{-1} for *E. coli* K-12, 200 μg ml^{-1} for EAEC) or kanamycin (50 μg ml^{-1}). Expression of genes from pETG20A and pRSF vectors was induced with 1 mM of isopropyl-β-D-thiogalactopyrannoside (IPTG, Eurobio) for 16 h.

Strain construction. Gene deletion into the enteroaggregative *E. coli* 17-2 strain was achieved by using a modified one-step inactivation procedure[41] as previously described[23] using plasmid pKOBEG[42]. Briefly, a kanamycin cassette was amplified from plasmid pKD4[41] using oligonucleotide pairs carrying 5′ 50-nucleotide extensions homologous to regions adjacent to the gene to be deleted. After electroporations of 600 ng of column-purified PCR product, kanamycin-resistant clones were selected and verified by colony-PCR. The kanamycin cassette was then excised using plasmid pCP20 (ref. 41). Gene deletions were confirmed by colony-PCR. The same procedure was used to introduce the *mCherry*- or *sfGFP*-coding sequences downstream from the start codon (vector pKD4-*sfGFP* or pKD4-*mCherry*) or the *mCherry*-coding sequence upstream from the stop codon (vector p*mCherry*-KD4). This procedure yields strains producing fusion proteins from their original chromosomal loci.

Plasmid construction. PCRs were performed using the Phusion DNA polymerase (Thermo Scientific). Restriction enzymes were purchased from New England Biolabs and used according to the manufacturer's instructions. Custom oligonucleotides were synthesized by Sigma Aldrich and are listed in Supplementary Table 1. Enteroaggregative *E. coli* 17-2 chromosomal DNA was used as a template for all PCRs. *E. coli* strain DH5α was used for cloning procedures. The pETG20A vector derivative encoding the periplasmic domain of the TssM periplasmic domain (TssMp, residues 386–1129) or the TssJ soluble domain have been previously described[25]. The fragment encoding TssM$_{26Ct}$ (residues 869–1107) was cloned into the pETG20A vector by restriction-free cloning[43]. The pRSF–TssJ^{S} intermediate plasmid was constructed by restriction cloning. Briefly, the sequence encoding the full-length *tssJ* gene (residues 1–178) was PCR-amplified using primers RSF-sJ-F and RSF-sJ-R. The PCR introduced a 5′ NdeI and a 3′ XhoI restriction site and a C-terminal streptavidin extension. The *tssJ* PCR product was sub-cloned into the pRSF-Duet (Novagen) MCS2 corresponding restriction sites. The pRSF–TssJ^{S}-^{F}L-^{H}M (encoding C-terminally StrepII-tagged TssJ, N-terminally Flag-tagged TssL and N-terminally 6×His-tagged TssM) was constructed by restriction-free cloning[43] as previously described[22]. Briefly, the sequence encoding the full-length *tssL* (residues 1–217) and full-length *tssM* (residues 1–1129) genes were PCR-amplified using the primer pairs RSF-fL-F/RSF-fL-R and RSF-hM-F/RSF-hM-R, respectively. The two PCR products (*tssL* and *tssM*) were synthesized with 30-base-pair overhangs, from both 5′ and 3′ ends, corresponding to the designed integration sites into the pRSF–TssJ^{S} plasmid. The double-stranded product of the first PCR was then used as oligonucleotides for a second PCR using the target vector as template. The introduction of the C1S mutation in TssJ was performed by QuikChange mutagenesis of the pRSF–TssJ^{S}-^{F}L-^{H}M plasmid using oligonucleotides Jcs-F and Jcs-R. Plasmid pIBA37-TssM was constructed by restriction-free cloning and cysteine derivatives were obtained by QuikChange mutagenesis using pIBA37-TssM-C757S mutant as template.

Hcp release assay. Cells producing Flag- or HA-tagged Hcp from plasmids pUCHcp$_{\mathrm{Flag}}$ or pOKHcp$_{\mathrm{HA}}$[22,23] were grown in SIM to an absorbance $A_{600\ \mathrm{nm}} \sim 0.8$. Supernatant and cell fractions were separated as previously described[23]. Briefly, 2×10^{9} cells were harvested and collected by centrifugation at 2,000*g* for 5 min. The supernatant fraction was then subjected to a second low-speed centrifugation and then at 16,000*g* for 15 min. The supernatant was filtered on sterile polyester membranes with a pore size of 0.2 μm (Membrex 25 PET, membraPure) before overnight precipitation with trichloroacetic acid 15% on ice. Cells and precipitated supernatants were resuspended in loading buffer and analysed by SDS–PAGE and immunoblotting with the anti-Flag or anti-HA antibody. As control for cell lysis, western blots were probed with antibodies raised against the periplasmic TolB protein. The assays were performed from three independent cultures, and a representative experiment is shown.

Interbacterial competition assay. The antibacterial growth competition assay was performed as described for the studies on the *Citrobacter rodentium* and EAEC Sci-2 T6SSs[21,44] with modifications. The wild-type *E. coli* strain W3110 bearing the Kan^{R} pUA66-*rrnB* plasmid[45] was used as prey in the competition assay. Attacker and prey cells were grown for 16 h in LB medium, then diluted in SIM to allow maximal expression of the *sci-1* gene cluster[40]. Once the culture reached $A_{600\ \mathrm{nm}} \sim 0.8$, the cells were harvested and normalized to $A_{600\ \mathrm{nm}} = 0.5$ in SIM. Attacker and prey cells were mixed to a 4:1 ratio and 20-μl drops of the mixture were spotted in triplicate onto a pre-warmed dry SIM agar plate supplemented or not with anhydrotetracyclin 0.02 μg ml^{-1}. After overnight incubation at 37 °C, the bacterial spots were then cut off, and cells were resuspended in SIM to $A_{600\ \mathrm{nm}} = 0.5$. Two hundred microlitres of serial dilutions were plated on kanamycin LB plates and the number of colonies was scored after overnight incubation at 37 °C. The assays were performed from at least three independent cultures, with technical triplicates and a representative technical triplicate shown.

Fluorescence microscopy, image treatment and statistical analyses. Fluorescence microscopy experiments were performed essentially as described[21,46]. Briefly, cells were grown overnight in LB medium and diluted to $A_{600\ \mathrm{nm}} \sim 0.04$ in SIM. Exponentially growing cells ($A_{600\ \mathrm{nm}} \sim 0.8$–1) were harvested, washed in phosphate buffered saline buffer (PBS), resuspended in PBS to $A_{600\ \mathrm{nm}} \sim 50$, spotted on a thin pad of 1.5% agarose in PBS, covered with a cover slip and incubated for 1 h at 37 °C before microscopy acquisition. For each experiment, ten independent fields were manually defined with a motorized stage (Prior Scientific) and stored (*x*, *y*, *z*, Perfect Focus System (PFS) offset) in our custom automation system designed for time-lapse experiments. Fluorescence and phase contrast micrographs were captured every 30 s using an automated and inverted epifluorescence microscope TE2000-E-PFS (Nikon) equipped with PFS. PFS automatically maintains focus so that the point of interest within a specimen is always kept in sharp focus at all times despite mechanical or thermal perturbations. Images were recorded with a CoolSNAP HQ 2 (Roper Scientific) and a ×100/1.4 DLL objective. The excitation light was emitted by a 120 W metal halide light. All fluorescence images were acquired with a minimal exposure time to reduce bleaching and phototoxicity effects. The sfGFP images were recorded by using the ET-GFP filter set (Chroma 49002) with an exposure time of 200–400 ms. The mCherry images were recorded by using the ET-mCherry filter set (Chroma 49008) using an exposure time of 100–200ms. Slight movements of the whole field during the time of the experiment were corrected by registering individual frames using StackReg and Image Stabilizer plugins for ImageJ. sfGFP and mCherry fluorescence channels were adjusted and merged using ImageJ (http://rsb.info.nih.gov/ij/). sfGFP fluorescence sets of data were treated to monitor foci detection. Noise and background were reduced using the 'Subtract Background' (20 pixels Rolling Ball) plugin from Fiji (Image J/National Institutes of Health). The sfGFP foci were automatically detected by simple image processing: (1) create a mask of cell surface and dilate, (2) detect the individual cells using the 'Analyse Particle' plugin of Fiji and (3) identify foci by the 'Find Maxima' process in Fiji. To avoid false positive results, each event was manually controlled in the original data. Microscopy analyses were performed at least six times, each in technical triplicate, and a representative experiment is shown. Box-and-whisker plots representing the number of detected foci for each strain were made with R software. To compare each population, *t*-tests were performed in R. Sub-pixel resolution tracking of fluorescent foci: Fluorescent foci were detected using a local and sub-pixel resolution maxima detection algorithm and tracked over time with a specifically developed plug-in for ImageJ. The *x* and *y* coordinates were obtained for each fluorescent focus on each frame. The mean square displacement was calculated as the distance of the foci from its location at t_0 at each time using R software and plotted over time. For each strain tested, the mean square displacement of at least ten individual focus trajectories was calculated.

Inner and outer membrane separation. Cells were broken using an Emulsiflex-C5 (Avestin) and the crude membrane fraction was isolated by ultracentrifugation at 100,000*g* for 45 min. Outer and inner membranes were separated by differential solubilization. Inner membranes were solubilized by 0.5% sodium *N*-lauroyl sarcosyl in 50 mM Tris-HCl pH 8.0 for 30 min at 20 °C. The insoluble material containing the outer membrane fraction was isolated by ultracentrifugation at 100,000*g* for 20 min. The outer membrane pellet was then solubilized in SDS-loading buffer. The assay was performed in triplicate, from three independent cultures and a representative experiment is shown.

TssJLM complex production and purification. The pRSF–TssJ^{S}-^{F}L-^{H}M plasmid was transformed into the *E. coli* BL21(DE3) expression strain (Invitrogen). Cells were grown at 37 °C in lysogeny broth (LB) to $A_{600\ \mathrm{nm}} \sim 0.7$ and the expression of the *tssJLM* genes was induced with 1.0 mM IPTG for 16 h at 16 °C. Cell pellets were resuspended in ice-cold 50 mM Tris-HCl pH 8.0, 50 mM NaCl, 1 mM EDTA and 1 mM TCEP, supplemented with 100 μg ml^{-1} of DNase I, 100 μg ml^{-1} of lysozyme and EDTA-free protease inhibitor (Roche). After sonication, $MgCl_2$ was added to the final concentration of 10 mM and the cell suspension was further broken using an Emulsiflex-C5 (Avestin). The broken cell suspension was clarified by centrifugation at 38,500*g* for 20 min. The membrane fraction was then collected by centrifugation at 98,000*g* for 45 min. Membranes were mechanically homogenized and solubilized in 50 mM Tris-HCl pH 8.0, 50 mM NaCl, 0.5% (w/v) *n*-dodecyl-β-D-maltopyranoside (DDM, Anatrace), 0.75% (w/v) decyl maltose neopentyl glycol (DM-NPG, Anatrace), 0.5% (w/v) digitonin (Sigma-Aldrich), 100 μM TCEP and 1 mM EDTA at 22 °C for 45 min. The suspension was clarified by centrifugation at 98,000*g* for 20 min. The supernatant was loaded onto a 5-ml StrepTrap HP (GE Healthcare) column and then washed with 50 mM Tris-HCl pH 8.0, 50 mM NaCl, 0.05% (w/v) DM-NPG (Affinity buffer) at 4 °C. The TssJLM core complex was eluted in the affinity buffer supplemented with 2.5 mM desthiobiotin (IBA) into a 5-ml HisTrap HP (GE Healthcare) column. The column was then washed in the affinity buffer supplemented with 20 mM imidazole and the TssJLM core complex was eluted in the same buffer supplemented with 500 mM imidazole. Peak fractions were pooled and loaded onto a Superose 6 10/300 column (GE Healthcare) equilibrated in 50 mM Tris-HCl pH 8.0, 50 mM NaCl, 0.025% (w/v) DM-NPG. The TssJLM core complex eluted as a single monodisperse peak close to the void volume of the column. The sample was used immediately for EM sample preparation.

Stoichiometry analyses. The purified TssJLM core membrane complex was diluted to a final concentration of 0.1 mg ml^{-1} and denaturated at 100 °C for 10 min after the addition of 1% sodium dodecyl sulfate. The denatured sample was incubated in presence of 40 μM of Alexa Fluor 532 C5-maleimide (Invitrogen) and 1 mM TCEP (Pierce) for 2 h at room temperature. The labelled proteins were separated by SDS–PAGE and protein-bound fluorescence was visualized and quantified using a Fujifilm FLA-3000 scanner. The assay was performed in triplicate, from three independent TssJLM complex preparations, and a representative experiment is shown. The quantification is expressed with the standard deviation from the three biological replicates.

EM and image processing. Determination of the TssJLM core membrane complex structure was achieved by negative-stain EM. Nine microlitres of suitably diluted (~0.01 mg ml^{-1}) TssJLM complex sample was spotted to glow-discharged carbon-coated copper grids (Agar Scientific). After 30 s of absorption, the sample was blotted, washed with three drops of water and then stained with 2% uranyl acetate. Images were recorded automatically using the EPU software on an FEG microscope operating at a voltage of 200 kV and a defocus range of 0.6–25 nm, using an FEI Falcon-II detector (Gatan) at a nominal magnification of 50,000, yielding a pixel size of 1.9 Å. A dose rate of 25 electrons per square ångström per second, and an exposure time of 1 s, were used. A total of 72,146 particles were automatically selected from 1,200 independent images and extracted within boxes of 280 pixels × 280 pixels using EMAN2/BOXER[47]. The defocus value was estimated and the contrast transfer function was corrected by phase flipping using EMAN2 (e2ctf). All 2D and 3D classifications and refinements were performed using RELION 1.3 (refs 48, 49). We used three rounds of reference-free 2D class averaging to clean up the automatically selected data set. Only highly populated classes displaying high-resolution features were conserved during this procedure and a final data set of 26,544 particles was assembled. An initial 3D model was generated in EMAN2 using 30 classes. Three-dimensional classification was then performed in Relion with five classes. The particles corresponding to the most populated class (~16,738) were used for refinement. Relion auto-refine procedure was used to obtain a final reconstruction at 11.56-Å resolution after masking and with C_5 symmetry imposed. Reported resolutions are based on the 'gold standard' Fourier shell correlation (FSC) 0.143 criterion, and FSC curves were corrected for the effects of a soft mask on them by using high-resolution noise substitution[50]. Before visualization, all density maps were corrected for the modulation transfer function of the detector and then sharpened by applying an ad hoc negative B-factor (−1,000). Local resolution variations were estimated using ResMap[51].

Three-dimensional maps display and analysis. Three-dimensional reconstructions were displayed and rendered in USCF Chimera segmented using the SEGGER module implemented in UCSF Chimera[52]. Segments corresponding to individual structural domains are represented in Fig. 3c–e. All other maps were left un-segmented. A volume/mass conversion of 0.81 Da Å^{-3} was used to estimate the volume occupied by TssM and TssL cytoplasmic domains.

Protein production and purification for SAXS and X-ray analyses. The periplasmic domain of the TssM protein, TssMp (residues 386–1129), was produced and purified as described previously[25]. The purified recombinant TssMp was digested with trypsin (at a 1,000:1 molecular ratio) at room temperature for 24 h. The reaction was quenched by the addition of 1 mM phenyl-methanesulfonyl fluoride (PMSF) and the insoluble TssMp fragments were discarded by centrifugation at 20,000*g* for 30 min. A proteolysis-resistant fragment of apparent size ~32 kDa (called hereafter $TssM_{32Ct}$) was further purified by consecutive ion-exchange (Mono Q 5/50 GL column, GE Healthcare) and size-exclusion (Superdex 75 16/600 HL column) chromatographies using an Äkta system (GE Healthcare). The purified fragment was subjected to N-terminal Edman sequencing. A PVDF membrane was rinsed three times with a water/ethanol mixture (10/90) and inserted in the A cartridge of a Procise 494A sequencer. After five cycles of Edman degradation, the sequence DYGSL was identified by mass spectrometry, indicating that cleavage after Arg834 generated a C-terminal fragment of theoretical mass 32,398 Da, in agreement with the 32-kDa band observed by SDS–PAGE analyses.

For production and purification of the $TssM_{26Ct}$ fragment (Thr869 to Glu1107), *E. coli* BL21(DE3) cells cultivated in the TB medium carrying plasmid pETG20A-$TssM_{26Ct}$ were grown to $A_{600\ nm}$ ~ 0.6 and the expression of $TssM_{26Ct}$ was induced by the addition of 0.5 mM IPTG for 16 h at 17 °C. Cells were collected by centrifugation at 10,000*g* at 4 °C for 15 min. The cell pellet was resuspended in lysis buffer and lysed by sonication. The lysate was clarified by centrifugation at 20,000*g* at 4 °C for 15 min, and the supernatant containing the Trx–His_6–tev–$TssM_{26ct}$ fusion protein was purified by consecutive Ni^{2+}-affinity and size-exclusion (Superdex 75 column) chromatographies on an Äkta purifier (GE Healthcare). The fractions containing the protein of interest were pooled and the 6×His-tagged TEV protease was added (5% w/w). The cleaved protein was purified using Ni^{2+} affinity, removing the Trx-His_6, followed by size-exclusion chromatography (Superdex 75 column) on an Äkta purifier (GE Healthcare). Over 100 mg of $TssM_{26ct}$ fragment were obtained per litre of culture. The purified protein was verified by mass spectrometry, before being concentrated up to 8.7 mg ml^{-1} in 20 mM Tris-HCl pH 8.0, NaCl 150 mM.

The production of nanobody nb25 and the formation of its complex with $TssM_{32Ct}$ have been described previously[31,32]. The production of unacylated TssJ was previously described[25]. The $TssM_{26Ct}$–TssJ complex was obtained by mixing $TssM_{26Ct}$ (8.7 mg ml^{-1}) with purified TssJ (30 mg ml^{-1}) in a 1:1.2 molecular ratio and the complex was then concentrated up to 15 mg ml^{-1} using centricon (cut-off of 10,000 Da) in 20 mM Tris-HCl pH 8.0, 150 mM NaCl.

SAXS and *ab initio* 3D shape reconstruction. SAXS analyses were performed at the ID29 beamline (European Synchrotron Radiation Facility, Grenoble, France) at a working energy of 12.5 keV ($\lambda = 0.931$ Å). Thirty microlitres of protein solution at 1.6 mg ml^{-1} in Tris-HCl 20 mM pH 8.0, NaCl 150 mM, were loaded by a robotic system into a 2-mm quartz capillary mounted in a vacuum. Ten independent 10-s exposures were collected on a Pilatus 6M-F detector placed at a distance of 2.85 m for each protein concentration. Individual frames were processed automatically and independently at the beamline by the data collection software (BsxCUBE), yielding radially averaged normalized intensities as a function of the momentum transfer *q*, with $q = 4\pi\sin(\theta)/\lambda$, where 2θ is the total scattering angle and λ is the X-ray wavelength. Data were collected in the range $q = 0.04$–6 nm^{-1}. The ten frames were combined to give the average scattering curve for each measurement. Data points affected by aggregation, possibly induced by radiation damage, were excluded. Scattering from the buffer alone was also measured before and after each sample analysis and the average of these two buffer measures was used for background subtraction using the program PRIMUS[53] from the ATSAS package[54]. PRIMUS was also used to perform Guinier analysis[55] of the low *q* data, which provides an estimate of the radius of gyration (R_g). Regularized indirect transforms of the scattering data were performed with the program GNOM[56] to obtain $P(r)$ functions of interatomic distances. The $P(r)$ function has a maximum at the most probable intermolecular distance and goes to zero at D_{max}, the maximum intramolecular distance. The values of D_{max} were chosen to fit with the experimental data and to have a positive $P(r)$ function. Three-dimensional bead models that fitted with the scattering data were built with the program DAMMIF[57]. Twenty independent DAMMIF runs were performed using the scattering profile of the TRX–His–TssJp and TRX–His–TssMp complexes, with data extending up to 0.35 nm^{-1}, using slow mode settings, assuming no symmetry and allowing for a maximum 500 steps to grant convergence. The models resulting from independent runs were superimposed using the DAMAVER suite[58]. This yielded an initial alignment of structures based on their axes of inertia followed by minimization of the normalized spatial discrepancy[59]. The normalized spatial discrepancy was therefore computed between a set of 20 independent reconstructions, with a range of normalized spatial discrepancies from 0.678 to 0.815. The aligned structures were then averaged, giving an effective occupancy to each voxel in the model, and filtered at half-maximal occupancy to produce models of the appropriate volume that were used for all subsequent analyses. All the models were similar in terms of agreement with the experimental data, as measured by DAMMIF χ parameter and the quality of the fit to the experimental curve. The DAMFILT average volume was used as the final model of the TRX–His–TssJ and TRX–His–TssMp complexes.

Ni-TNA-Nanogold labelling. The TssJLM complex was spotted onto a glow-discharged carbon coated grid (CF-400, Electron Microscopy Sciences). After 1 min, excess liquid was blotted, and the grid was washed on a drop of cold purification buffer (50 mM Tris pH 8, 50 mM NaCl, 0.025% (w/v) DM-NPG) containing 50 mM imidazole, quickly blotted and deposited on a second drop of the same buffer in the presence of 5 nM Ni-TNA-Nanogold beads (Nanoprobes). After 2 min, the grid was rinsed sequentially for 20 s with one drop of purification buffer, one drop of the same buffer without detergent and three drops of 2% uranyl acetate. Images were collected on an FEI Tecnai F20 FEG microscope operating at a voltage of 200 kV, equipped with a direct electron detector (Falcon II). Particles were selected manually using EMAN2. The assay was performed at least in triplicate, from independent TssJLM complex preparations, and representative particles are shown.

Anti-Strep labelling. The TssJLM complex was mixed with monoclonal anti-Strep antibodies (Sigma) at a ratio of complex:antibody of 2:1. The mixture was incubated at 4 °C for 30 min and the labelled complex was isolated by gel filtration. The sample was analysed by negative-stain EM as described above for negative-stain EM of the unlabelled TssJLM complex. The assay was performed at least in triplicate, from independent TssJLM complex preparations, and representative particles are shown.

Crystallization and structure determination. The crystallization of the TssM$_{32Ct}$–nb25 complex has been described previously[31]. For TssM$_{26Ct}$ alone, several kits were used for crystallization screening, including STURA, WIZARD, MDL, INDEX and PEGs. A hit was observed in the PEGs kit, within a well containing 0.2 M zinc acetate and 20% (m/v) PEG3350. Crystal optimization was performed by varying PEG3350 amount in the 15–25% range in 0.1 M sodium acetate and 0.2 M $ZnCl_2$ at pH varying between 3.8 and 5.5. Crystals appeared after few days in 20% PEG3350, 0.1 M sodium acetate pH 4 and 0.2 M $ZnCl_2$. Crystals were tested at the European Synchrotron Radiation Facility (ESRF) ID23-1 beamline after cryo-cooling in the crystallization liquor supplemented with 12.5% propylene glycol.

The TssM$_{26Ct}$–TssJ complex was screened for crystallization using the PEGs and PACT1 kits. Hits were observed in PACT1. All contained Zn^{2+}: 0.01 M zinc chloride, 0.1 M sodium acetate pH 5, 20% (w/v) PEG 6000; or 0.01 M zinc chloride, 0.1 M MES pH 6, 20% (w/v) PEG 6000. Crystal optimization was performed by using PEG6000 in the 10–20% range in 0.1 M sodium acetate/MES pH 4.75–6, and $ZnCl_2$ at 0, 0.01, 0.05 and 0.2 M. No crystals were obtained in conditions without $ZnCl_2$ or containing 0.2 M $ZnCl_2$. By contrast, well-shaped crystals appeared in 50 mM $ZnCl_2$, 15% PEG6000 and 0.1 sodium acetate pH 4.75. Crystals were cryo-cooled with polypropylene glycol 12.5% but diffracted to only ~4.0 Å at the Soleil Proxima 2 beamline (Saint Aubin, France). Further crystals were obtained in LIMBRO plates. Large crystals were obtained by mixing 6 μl of protein and 2 μl of well solution in 50 mM $ZnCl_2$, 15% PEG6000, 94 mM sodium acetate pH 4.75 and 6 mM MES pH 6. Crystals were dipped in polypropylene glycol for ~20 s and exposed at the ESRF ID23-1 beamline.

Data collection was performed at 100 K at the Soleil Proxima 1 beamline (Saint Aubin, France) for TssM$_{32Ct}$–nb25 and at the ID23-1 beamline (ESRF synchrotron, Grenoble, France) for TssM$_{26Ct}$ alone and for the TssM$_{26Ct}$–TssJ complex. Data were processed by the XDS[60] package and scaled with XSCALE (Extended Data Table 1).

The structure of the TssM$_{32Ct}$–nb25 complex was determined by molecular replacement with Molrep[61] using the previously determined structure of nb25 (Protein Data Bank accession number 4QGY)[32]. After refining the positions of the two nb25 molecules in the asymmetric unit by rigid body refinement with AutoBuster[62], an electron density map was calculated at 1.92-Å resolution. Features such as α-helices were easily identified, making it possible to trace manually the model of TssM$_{32Ct}$ using COOT[63], alternated with cycles of refinement with AutoBuster[62], with non-crystallographic symmetry restraints, and translation, rotation and screw-rotation (TLS) group refinement[64], features used in all refinement procedures described below. The final structure at 1.92-Å resolution had R_{work}/R_{free} values of 18.4/21.0%, 96.3% of the residues in the preferred area of the Ramachandran plot and no outliers (Extended Data Table 1).

The structure of TssM$_{26Ct}$ alone was solved by molecular replacement with Molrep[61] using the refined model of TssM$_{32Ct}$ from the TssM$_{32Ct}$–nb25 complex. The initial structure model was improved through iterative refinement with AutoBuster[62] and manual refitting with COOT[63]. The final structure at 1.51-Å resolution had R_{work}/R_{free} values of 19.2/20.2%, 97.4% of the residues in the preferred area of the Ramachandran plot and no outliers (Extended Data Table 1).

The structure of the TssM$_{26Ct}$–TssJ complex was solved by molecular replacement with Molrep[61] using the refined structure of TssM$_{26ct}$ and the previously determined TssJ structure (Protein Data Bank 3RX9)[25] in which all the other conformations were removed. A first round of rigid body refinement and four cycles of Phenix[65] cartesian-simulated annealing were performed. The resulting model was improved through iterative refinement with AutoBuster[62] and manual refitting with COOT[63]. The final structure at 2.24-Å resolution had R_{work}/R_{free} values of 20.0/22.3%, 96.9% of the residues in the preferred area of the Ramachandran plot and four outlier residues in very poorly defined loops (Extended Data Table 1). The TssM$_{32Ct}$–nb25, TssM$_{26Ct}$ and TssM$_{26Ct}$–TssJ structures form similar homodimers in the asymmetric unit. However, as reported by PISA[66], and the known topologies of TssM and TssJ[23,24], these dimers are not biologically relevant. Molecular contacts were analysed by the PISA server[33] and figures were prepared with Chimera[52] and Pymol[67].

The crystal structures of the TssM$_{32Ct}$–nb25 complex, and of the TssM$_{26Ct}$ fragment and TssM$_{26Ct}$–TssJ complexes, have been deposited in the Protein Data Bank under accession numbers 4Y7M, 4Y7L and 4Y7O respectively.

Docking TssM$_{26ct}$–TssJ structure and TssL$_{cyto}$. The crystal structures of the TssM$_{26ct–TssJ}$ complex and of the TssL cytoplasmic domain (Protein Data Bank 3U66)[28] were docked automatically using Chimera[52] after map segmentation.

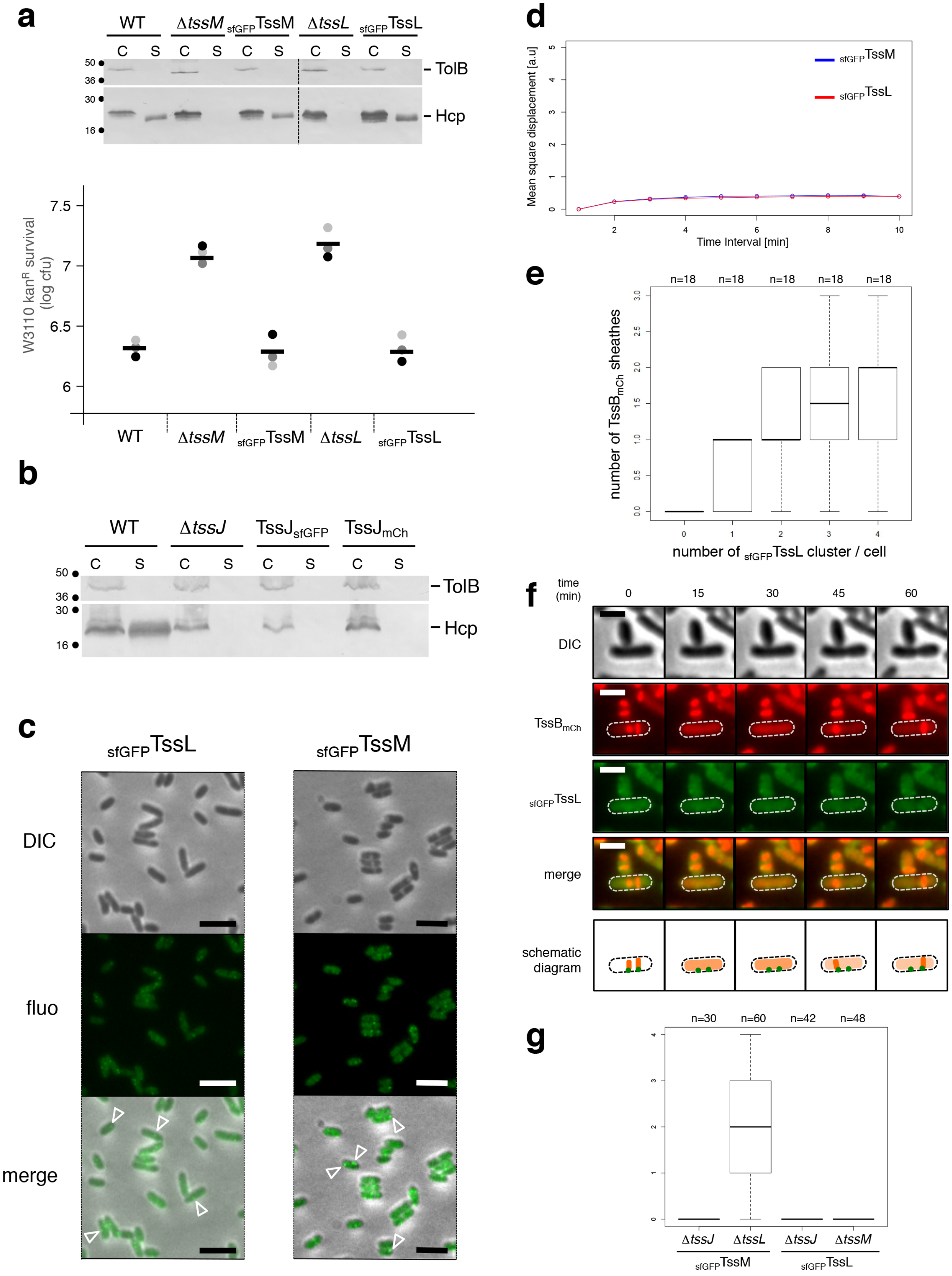
Refinement of docked TssM$_{26Ct}$-J pentamer in the EM density map. The atomic model of the docked TssM$_{26Ct}$-J structure was refined in the EM density with RSRef[68]. First, missing side-chain and polar hydrogen atoms were added with Modeller[69]. The structure was minimized using 2,000 steps of least-squares conjugate gradient refinement in the presence of distance restraints for hydrogen bonds and backbone dihedral angle restraints to maintain secondary structures. The minimization was performed with the real-space objective function calculated by RSRef in CNS[70]. The C_5 symmetry was enforced by strict non-crystallographic symmetry restraints. The total energy included internal parameters (bond length, bond angle, improper and dihedral angles) and non-bonded interactions with full Van der Waals and electrostatic potentials using a 7.5 Å cutoff. The final correlation coefficient between the EM reconstruction and the refined TssM$_{26Ct}$-J atomic model was 0.929 (as calculated by RSRef), whereas it was 0.706 before minimization.

Modelling of TssM$_{26Ct}$-J decamer. The atomic position of a TssM$_{26Ct}$-J protomer from the outer ring of the pentamer served as the starting structure to generate a TssM$_{26Ct}$-J decamer model with cyclic ten-fold symmetry using CNS[70]. The symmetry was enforced by strict non-crystallographic symmetry restraints (rotations of 36° around the symmetry axis). First, 5,000 steps of rigid body minimization were performed including only inter-protomer energetic contributions (full Van der Waals and electrostatic potentials). After a short all-atom minimization (300 steps), 1.5 ps of molecular dynamics simulation at 1,000 K was performed, followed by 300 steps of minimization and 10 ps of molecular dynamics simulation at 200 K. Minimizations and molecular dynamics simulations were realized with both intra-protomer and inter-protomer energetic contributions activated, and the backbone conformation of the protomer was restrained with harmonic constraints.

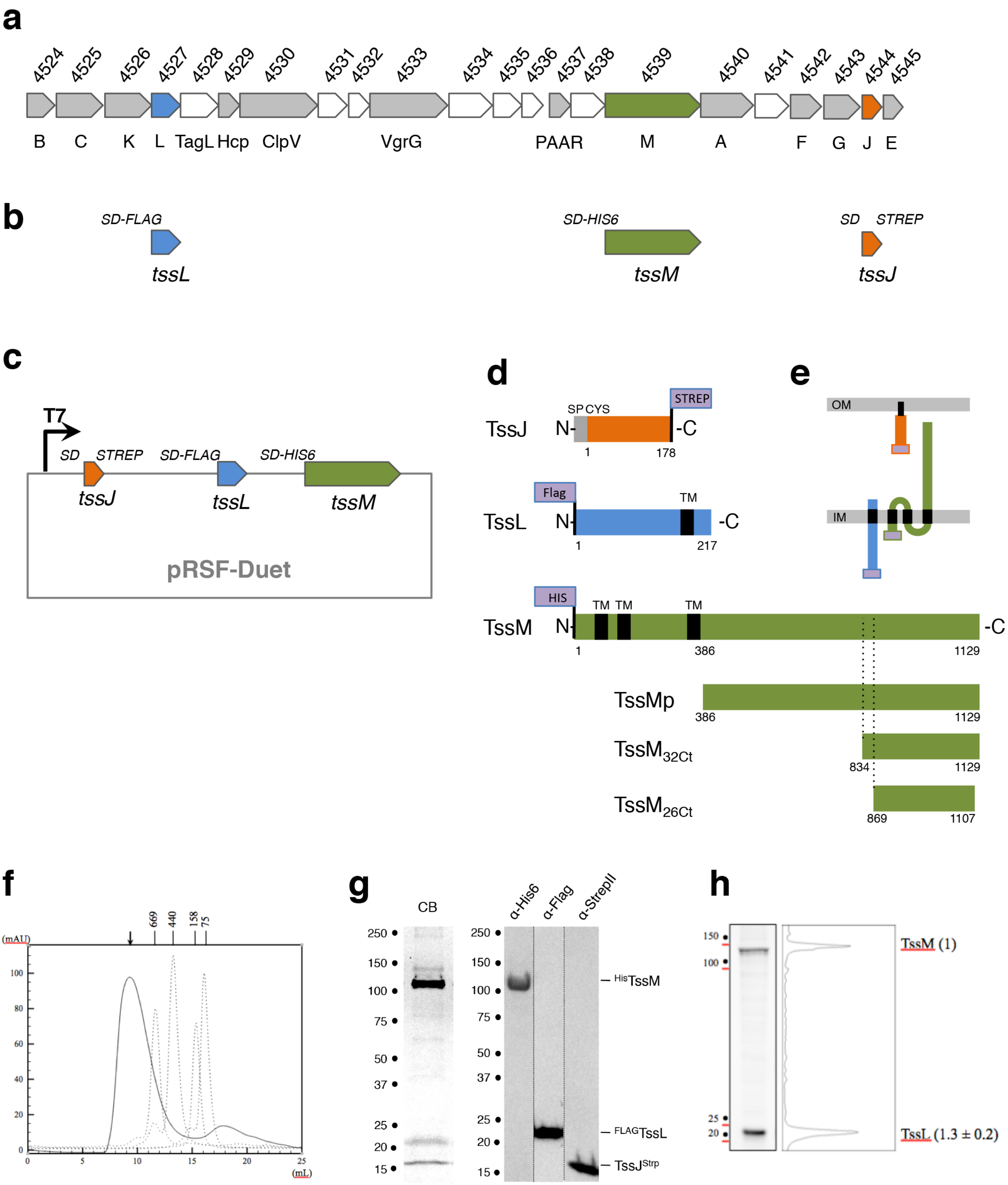
Substituted cysteine accessibility method. Cysteine accessibility experiments were performed on whole cells, mainly as described[71,72] with modifications. A 20-ml culture of wild-type or Δ*tssBC* strains producing a periplasmic cysteine-less TssM (Cys727-to-Ser) or derivatives bearing cysteine substitutions were induced for *tssM* gene expression with 0.05 μg ml^{-1} anhydrotetracyclin (AHT) for 1 h. Cells were harvested and resuspended in buffer A (100 mM Hepes (pH 7.5), 150 mM NaCl, 25 mM $MgCl_2$) to a final $A_{600\ nm}$ of 12 in 500 μl of buffer A. Bovine serum albumin (BSA)-coupled maleimide (Sigma-Aldrich) was added to a final concentration of 100 μM (from a 20 mM stock freshly dissolved in DMSO) and the cells were incubated for 30 min at 25 °C. β-Mercaptoethanol (20 mM final concentration) was added to quench the biotinylation reaction, then cells were washed twice in buffer A and resuspended in buffer A containing *N*-ethyl maleimide (final concentration 5 mM) to block all free sulfhydryl residues. After incubation for 20 min at 25 °C, cells were disrupted by sonication. Membranes recovered by ultracentrifugation at 100,000*g* for 40 min were resuspended in Laemmli buffer before SDS–PAGE analysis and immunodetection with anti-Flag antibodies (to detect the TssM proteins). Controls were performed by labelling total membranes from the same samples instead of whole cells. The assay was performed in triplicate, from three independent cultures, and a representative experiment is shown.

SDS–PAGE, protein transfer, immunostaining and antibodies. SDS–PAGE was performed on Bio-Rad Mini-PROTEAN systems using standard protocols. For immunostaining, proteins were transferred onto 0.2-μm nitrocellulose membranes (Amersham Protran). Immunoblots were probed with primary antibodies and goat secondary antibodies coupled to alkaline phosphatase, and developed in alkaline buffer in presence of 5-bromo-4-chloro-3-indolylphosphate and nitroblue tetrazolium. The anti-TolB was from our laboratory collection, whereas the anti-HA (3F10 clone, Roche), anti-Flag (M2 clone, Sigma Aldrich), anti-StrepII (Sigma Aldrich), anti-5his (Sigma Aldrich) monoclonal antibodies and alkaline-phosphatase-conjugated goat anti-rabbit or mouse secondary antibodies (Millipore) were purchased as indicated.

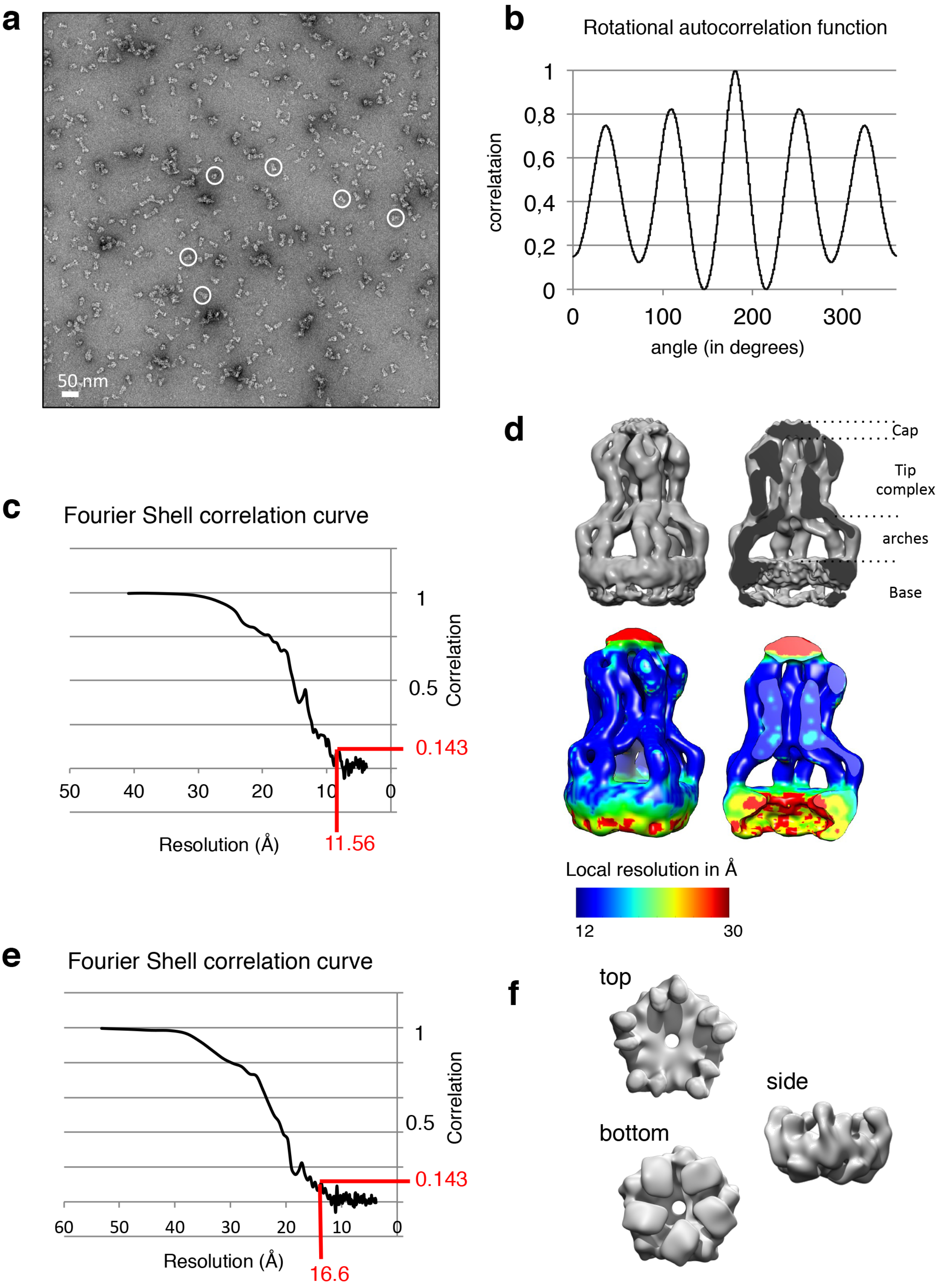
40. Brunet, Y. R., Bernard, C. S., Gavioli, M., Lloubès, R. & Cascales, E. An epigenetic switch involving overlapping fur and DNA methylation optimizes expression of a type VI secretion gene cluster. *PLoS Genet.* **7,** e1002205 (2011).
41. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97,** 6640–6645 (2000).
42. Chaveroche, M. K., Ghigo, J. M. & d'Enfert, C. A rapid method for efficient gene replacement in the filamentous fungus *Aspergillus nidulans. Nucleic Acids Res.* **28,** e97 (2000).
43. van den Ent, F. & Löwe, J. RF cloning: a restriction-free method for inserting target genes into plasmids. *J. Biochem. Biophys. Methods* **67,** 67–74 (2006).
44. Gueguen, E. & Cascales, E. Promoter swapping unveils the role of the *Citrobacter rodentium* CTS1 type VI secretion system in interbacterial competition. *Appl. Environ. Microbiol.* **79,** 32–38 (2013).
45. Zaslaver, A. *et al.* A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli. Nature Methods* **3,** 623–628 (2006).
46. Zoued, A. *et al.* TssK is a trimeric cytoplasmic protein interacting with components of both phage-like and membrane anchoring complexes of the type VI secretion system. *J. Biol. Chem.* **288,** 27031–27041 (2013).
47. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157,** 38–46 (2007).
48. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180,** 519–530 (2012).
49. Scheres, S. H. Semi-automated selection of cryo-EM particles in RELION-1.3. *J. Struct. Biol.* **189,** 114–122 (2015).
50. Chen, S. *et al.* High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy* **135,** 24–35 (2013).
51. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods* **11,** 63–65 (2014).
52. Pettersen, E. F. *et al.* UCSF Chimera – a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
53. Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. & Svergun, D. I. PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *J. Appl. Crystallogr.* **36,** 1277–1282 (2003).
54. Konarev, P. V., Petoukhov, M. V., Volkov, V. V. & Svergun, D. I. ATSAS 2.1, a program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* **39,** 277–286 (2006).
55. Guinier, A. La diffraction des rayons X aux très petits angles; application à l'étude de phénomènes ultramicroscopiques. *Ann. Phys. (Paris)* **12,** 161–237 (1939).
56. Svergun, D. I. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J. Appl. Crystallogr.* **25,** 495–503 (1992).
57. Franke, D. & Svergun, D. I. DAMMIF, a program for rapid *ab-initio* shape determination in small-angle scattering. *J. Appl. Crystallogr.* **42,** 342–346 (2009).
58. Volkov, V. V. & Svergun, D. I. Uniqueness of *ab initio* shape determination in small-angle scattering. *J. Appl. Crystallogr.* **36,** 860–864 (2003).
59. Kozin, M. B. & Svergun, D. I. Automated matching of high- and low-resolution structural models. *J. Appl. Crystallogr.* **34,** 33–41 (2001).
60. Kabsch, W. XDS. *Acta Crystallogr. D* **66,** 125–132 (2010).
61. Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr. D* **66,** 22–25 (2010).
62. Blanc, E. *et al.* Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT. *Acta Crystallogr. D* **60,** 2210–2221 (2004).
63. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
64. Winn, M. D., Murshudov, G. N. & Papiz, M. Z. Macromolecular TLS refinement in REFMAC at moderate resolutions. *Methods Enzymol.* **374,** 300–321 (2003).
65. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
66. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372,** 774–797 (2007).
67. The PyMOL Molecular Graphics System. v.1.5.0.4 (Schrödinger, LLC, 2014).
68. Chapman, M. S., Trzynka, A. & Chapman, B. K. Atomic modeling of cryo-electron microscopy reconstructions – joint refinement of model and imaging parameters. *J. Struct. Biol.* **182,** 10–21 (2013).
69. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234,** 779–815 (1993).
70. Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nature Protocols* **2,** 2728–2733 (2007).
71. Bogdanov, M., Zhang, W., Xie, J. & Dowhan, W. Transmembrane protein topology mapping by the substituted cysteine accessibility method (SCAM™): application to lipid-specific membrane protein topogenesis. *Methods* **36,** 148–171 (2005).
72. Goemaere, E. L., Devert, A., Lloubès, R. & Cascales, E. Movements of the TolR C-terminal domain depend on TolQR ionizable key residues and regulate activity of the Tol complex. *J. Biol. Chem.* **282,** 17749–17757 (2007).
73. Du, D. *et al.* Structure of the AcrAB–TolC multidrug efflux pump. *Nature* **509,** 512–515 (2014).
74. Hodgkinson, J. L. *et al.* Three-dimensional reconstruction of the *Shigella* T3SS transmembrane regions reveals 12-fold symmetry and novel features throughout. *Nature Struct. Mol. Biol.* **16,** 477–485 (2009).
75. Low, H. H. *et al.* Structure of a type IV secretion system. *Nature* **508,** 550–553 (2014).


a
WT
ΔtssM
sfGFPTssM
ΔtssL
sfGFPTssL
C S C S C S C S C S
50
36
30
16
TolB
Hcp
W3110 kanR survival (log cfu)
7.5
7
6.5
6
WT
ΔtssM
sfGFPTssM
ΔtssL
sfGFPTssL
b
WT
ΔtssJ
TssJsfGFP
TssJmCh
C S C S C S C S
50
36
30
16
TolB
Hcp
c
sfGFPTssL
sfGFPTssM
DIC
fluo
merge
d
Mean square displacement [a.u]
0 1 2 3 4 5
sfGFPTssM
sfGFPTssL
2 4 6 8 10
Time Interval [min]
e
n=18 n=18 n=18 n=18 n=18
number of TssBmCh sheathes
0.0 0.5 1.0 1.5 2.0 2.5 3.0
0 1 2 3 4
number of sfGFPTssL cluster / cell
f
time (min)
0 15 30 45 60
DIC
TssBmCh
sfGFPTssL
merge
schematic diagram
g
n=30 n=60 n=42 n=48
0 1 2 3 4
ΔtssJ ΔtssL ΔtssJ ΔtssM
sfGFPTssM
sfGFPTssL

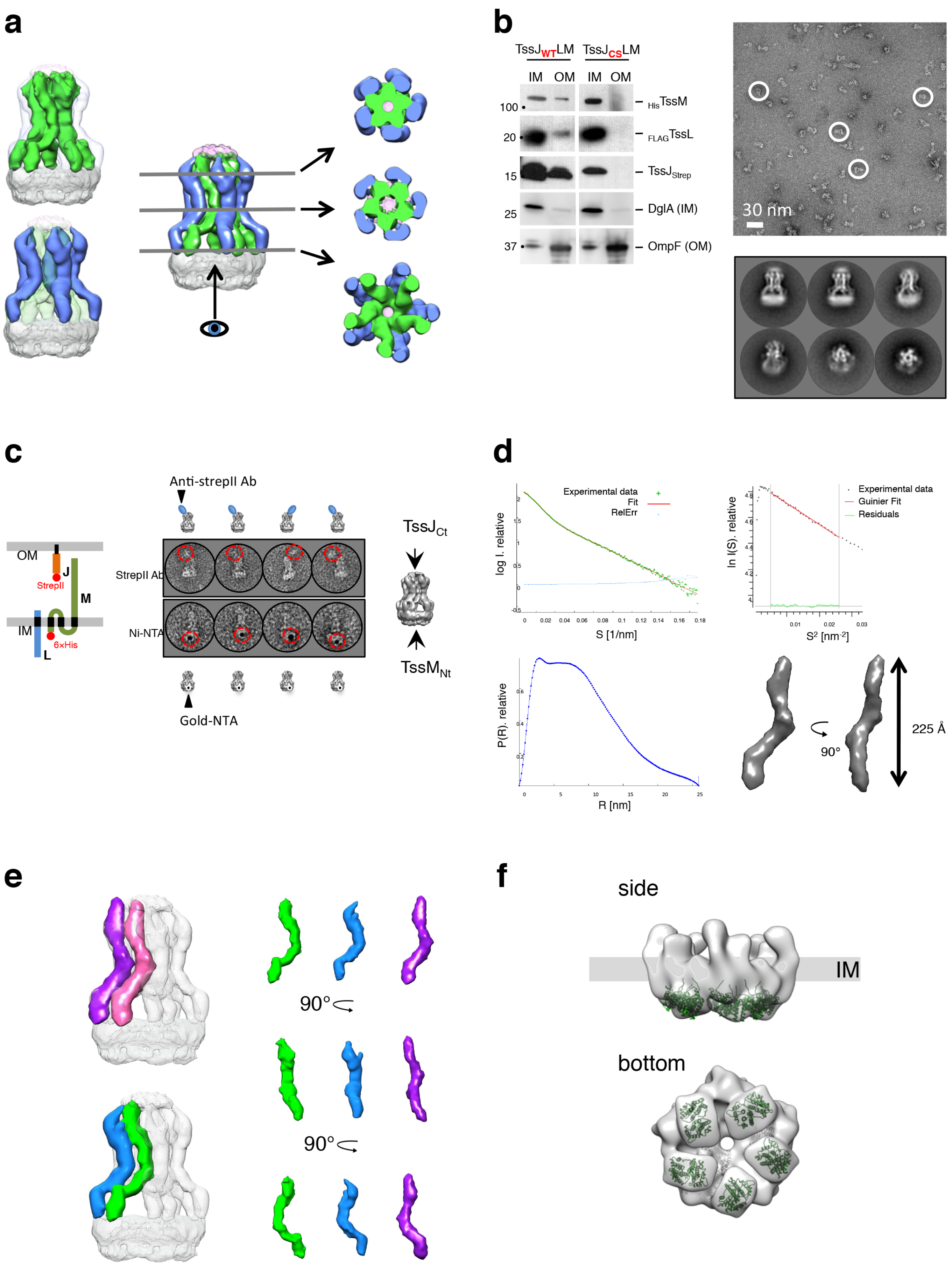
Extended Data Figure 1 | Functional and dynamic properties of fluorescently labelled Tss proteins. **a**, GFP–TssM and GFP–TssL fusion proteins are functional. Top: Hcp release assay. Hcp release was assessed by separating whole cells (C) and supernatant (S) fractions from the indicated strains. A total of 1×10^9 cells and the TCA-precipitated material from the supernatant of 2×10^9 cells were analysed by western blot using anti-Flag monoclonal antibody (lower panel) and anti-TolB polyclonal antibodies as a lysis control (upper panel). The molecular mass markers (in kilodaltons) are indicated on the left. Bottom: anti-bacterial assay. The anti-bacterial activity was assessed by mixing kanamycin-resistant prey *E. coli* K-12 cells with the indicated attacker cells for 16 h at 37 °C in SIM. The number of recovered *E. coli* prey cells is indicated in the graph (as log of colony-forming units (c.f.u.)). The circles indicate values from three independent assays, and the average is indicated by the bar. **b**, TssJ–sfGFP and TssJ–mCh fusion proteins are non-functional. Hcp release was assessed by separating whole cells (C) and supernatant (S) fractions from the indicated strains. A total of 1×10^9 cells and the TCA-precipitated material from the supernatant of 2×10^9 cells were analysed by western blot using anti-Flag monoclonal antibody (lower panel) and anti-TolB polyclonal antibodies as a lysis control (upper panel). The molecular mass markers (in kilodaltons) are indicated on the left. **c**, sfGFP–TssM and sfGFP–TssL cluster in foci. Large fields of fluorescence microscopy recordings showing localization of the sfGFP–TssL (left) and sfGFP–TssM (right) fusion proteins. The positions of selected foci are indicated by arrowheads. Scale bars, 5 μm. **d**, sfGFP–TssM and sfGFP–TssL foci are stable and static. Mean square displacement (in arbitrary units (a.u.)) of sfGFP–TssM (blue line) and sfGFP–TssL (red line) clusters were measured by sub-pixel tracking of fluorescent foci and plotted over time (in minutes). **e**, The TssBC sheath tubular structures assemble on TssJLM membrane complexes. Statistical analyses reporting the average number of sheath per cell compared with the number of membrane complexes per cell, highlighting the observation that the number of membrane complexes is at least equal to the number of sheathes. Lower and upper boundaries of the boxes correspond to the 25% and 75% percentiles respectively. Black bold horizontal bar, median values for each strain; whiskers, 10% and 90% percentiles; *n* indicates the number of cells studied per strain. **f**, Long-term fluorescence microscopy recordings. Time-lapse fluorescence microscopy recordings showing localization and dynamics of the sfGFP–TssL and TssB–mCherry fusion proteins. Individual images were taken every 15 min. Assembly/contraction of the sheath and TssL localization events are schematized in the lowest panel. Scale bars, 1 μm. **g**, Statistical analysis of sfGFP–TssM and sfGFP–TssL localization in various *tss* backgrounds. Shown are box-and-whisker plots of the measured number of sfGFP–TssM and sfGFP–TssL foci per cell for each indicated strain with the lower and upper boundaries of the boxes corresponding to the 25% and 75% percentiles respectively (horizontal bar, the median values for each strain; whiskers, the 10% and 90% percentiles); *n* indicates the number of cells studied per strain.


a
4524
4525
4526
4527
4528
4529
4530
4531
4532
4533
4534
4535
4536
4537
4538
4539
4540
4541
4542
4543
4544
4545
B
C
K
L
TagL
Hcp
ClpV
VgrG
PAAR
M
A
F
G
J
E
b
SD-FLAG
tssL
SD-HIS6
tssM
SD
STREP
tssJ
c
T7
SD
STREP
SD-FLAG
SD-HIS6
tssJ
tssL
tssM
pRSF-Duet
d
TssJ
N-
SP
CYS
STREP
-C
1
178
TssL
N-
Flag
TM
-C
1
217
TssM
N-
HIS
TM
TM
TM
-C
1
386
1129
TssMp
386
1129
TssM32Ct
834
1129
TssM26Ct
869
1107
e
OM
IM
f
(mAU)
669
440
158
75
100
80
60
40
20
0
0
5
10
15
20
25
(mL)
g
CB
250
150
100
75
50
37
25
20
15
α-His6
α-Flag
α-StrepII
250
150
100
75
50
37
25
20
15
HisTssM
FLAGTssL
TssJStrp
h
150
100
25
20
TssM (1)
TssL (1.3 ± 0.2)

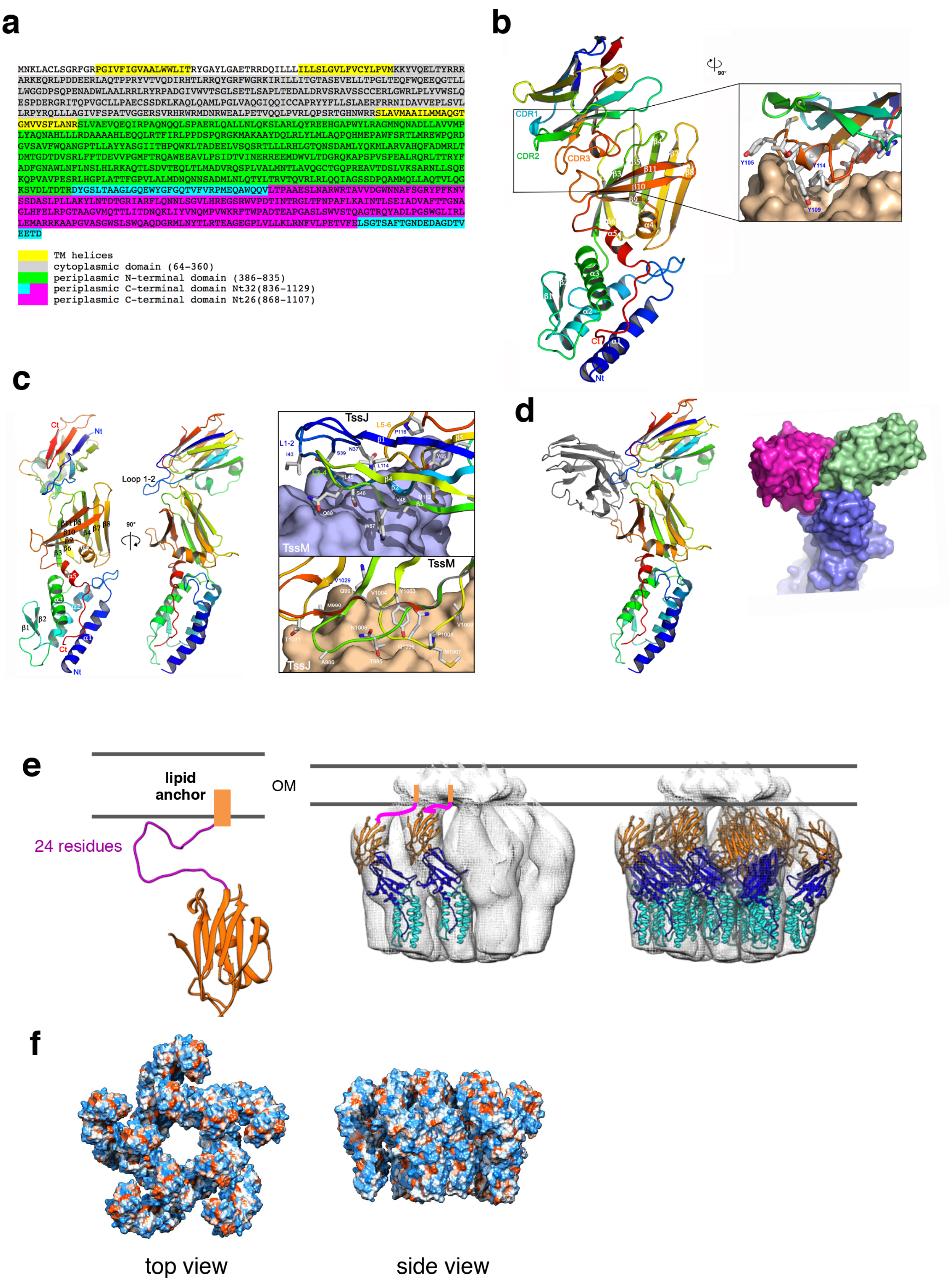
Extended Data Figure 2 | Expression and purification of the T6SS membrane core complex. **a**–**e**, T6SS operon genomic organization and constructs used for *in vitro* analyses. **a**, Schematic representation of the T6SS *sci-1* gene cluster from entero aggregative *E. coli*. The numbers on top refer to the gene locus tag (EC042_XXXX). Genes encoding core components (identified by their names on bottom, for example, 'B' refers to the *tssB* gene) are coloured grey. Genes of unknown function are coloured white. The three genes used to reconstitute the core membrane complex are coloured orange (*tssJ*), blue (*tssL*) and green (*tssM*). **b**, Schematic representation of the engineered constructs: the *tssJ*, *tssL* and *tssM* genes were amplified with an additional Shine Dalgarno (SD) sequence and 3′ StrepII, 5′ Flag and 5′ 6×His tags respectively. These three fragments were cloned into the pRSF-Duet vector (**c**). This construct allows the production of the C-terminally StrepII-tagged TssJ outer membrane (OM) lipoprotein and N-terminally Flag-tagged TssL and 6×His-tagged TssM inner-membrane (IM) proteins (**d**, **e**). The proteins are schematized and their boundaries and principal characteristics (TM, transmembrane segments; SP, signal peptide; CYS, acylated cysteine) are indicated (**d**) and their topologies are shown (**e**). The additional TssM constructs (TssMp, TssM$_{32Ct}$ and TssM$_{26Ct}$) used for SAXS or X-ray analyses are shown at the bottom. **f**–**h**, Purification and biochemical characterization of the T6SS membrane core complex. **f**, Analytical size-exclusion chromatography analysis of the purified TssJLM complex (continuous line) on a Superose 6 column, calibrated with 75-, 158-, 440- and 660-kDa molecular mass markers (dotted lines). The molecular mass of each marker (in kilodaltons) is indicated on the top of the corresponding peak. An arrow indicates the position of the peak fraction corresponding to the TssJLM complex. **g**, SDS–PAGE of the purified TssJLM complex analysed by Coomassie staining (CB) or immunoblotting using anti-His (α-His), -Flag (α-Flag) and -StrepII (α-STREP) antibodies. **h**, Left: cysteine labelling of the purified TssJLM complex in reducing and denaturing conditions as described in Methods. The total number of cysteine residues was nine for TssM, five for TssL and none for TssJ (the N-terminal cysteine is acylated). Right: the relative amount of TssL compared with TssM (densitometry relative to the number of free cysteine residues, fixed at 1 for TssM).


a
50 nm
b
Rotational autocorrelation function
correlataion
1
0,8
0,6
0,4
0,2
0
0
100
200
300
angle (in degrees)
c
Fourier Shell correlation curve
1
0.5
Correlation
0.143
50
40
30
20
10
0
Resolution (Å)
11.56
d
Cap
Tip complex
arches
Base
Local resolution in Å
12
30
e
Fourier Shell correlation curve
1
0.5
Correlation
0.143
60
50
40
30
20
10
0
Resolution (Å)
16.6
f
top
side
bottom

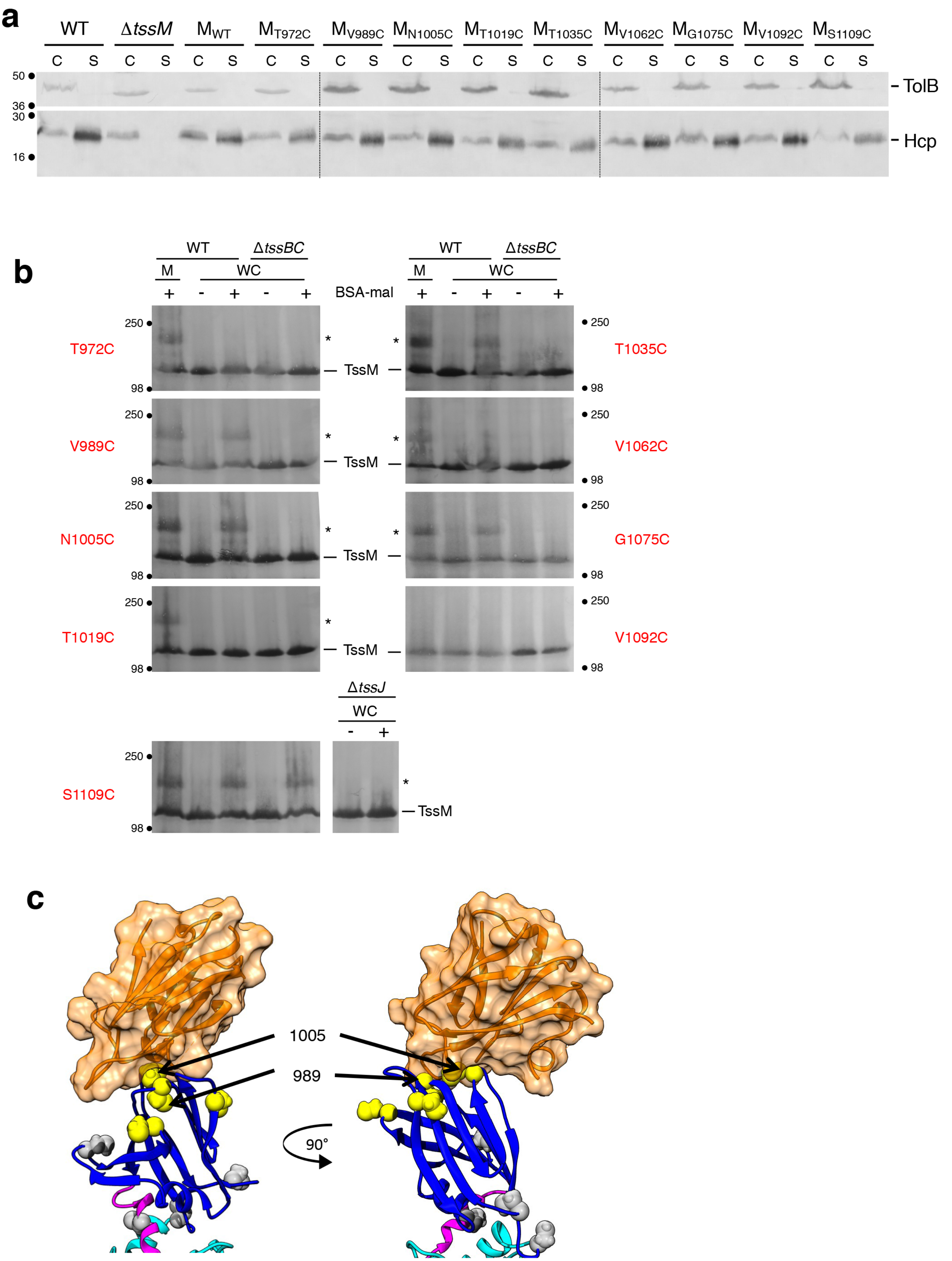
Extended Data Figure 3 | Architecture of the T6SS membrane core complex. **a**, Negative-stain EM of the EAEC TssJLM complex. Representative micrograph of the data set used for image processing. Isolated TssJLM complexes were clearly visible (white circles). **b**, Plot of the rotational autocorrelation function for a representative class average of an end-view. **c**, FSC curve of the TssJLM reconstruction. The 'gold standard' FSC curve was calculated in Relion using the masked reconstruction of the TssJLM complex. The resolution at 0.143 correlation was 11.56 Å. **d**, Top: side and corresponding cut-away views of the 3D reconstruction for the whole TssJLM complex. Bottom: local resolution as calculated by Resmap. The TssJLM volume (left reconstruction, side view; right reconstruction, cut-away view) is coloured according to the local resolution from high resolution (~12 Å) in blue to low resolution (>30 Å) in red. **e**, FSC curve of the TssJLM base. The 'gold standard' FSC curve was calculated in Relion using the unmasked reconstruction of the TssJLM base. The resolution at 0.143 correlation was 16.6 Å. **f**, Top, side and bottom views of the 3D reconstruction after specific refinement of the base.


a
b
TssJ$_{WT}$LM
TssJ$_{CS}$LM
IM
OM
IM
OM
100
20
15
25
37
$_{His}$TssM
$_{FLAG}$TssL
TssJ$_{Strep}$
DglA (IM)
OmpF (OM)
30 nm
c
Anti-strepII Ab
OM
J
StrepII
M
IM
6xHis
L
StrepII Ab
Ni-NTA
Gold-NTA
TssJ$_{Ct}$
TssM$_{Nt}$
d
Experimental data
Fit
RelErr
log I. relative
S [1/nm]
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
-0.5
0
0.5
1
1.5
2
Experimental data
Guinier Fit
Residuals
ln I(S). relative
S^2 [nm^{-2}]
0.01
0.02
0.03
4
4.2
4.4
4.6
4.8
P(R). relative
R [nm]
0
5
10
15
20
25
0.2
0.4
0.6
90°
225 Å
e
90°
90°
f
side
IM
bottom

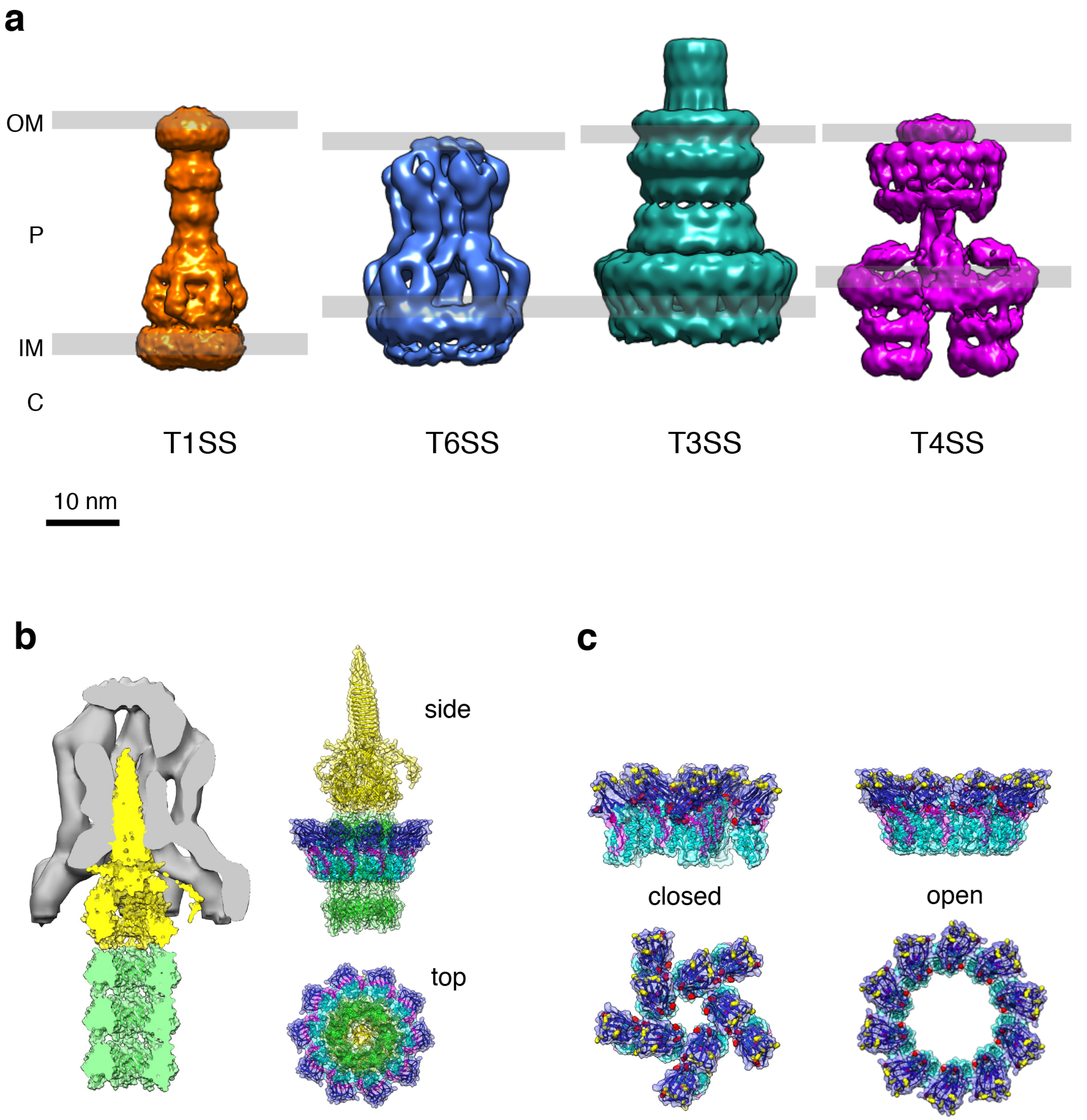
Extended Data Figure 4 | Structural analysis and segmentation. **a**, Segmentation of the TssJLM reconstruction. Left: above the base, ten equivalent densities could be defined by segmentation. They are arranged in two concentric rings. The internal ring is represented in green in the top panel and the external ring is represented in blue in the bottom panel. Right: cut-out views of the complex showing the arrangement of the two concentric rings at different levels (grey lines) along the periplasmic portion of the TssJLM complex. The cut-out views are seen from the bottom of the complex. **b**, Requirement of TssJ lipidation for complex assembly and insertion into the outer membrane. Left: membrane fractionation by differential solubilization followed by immunoblot analysis. Total membrane extracts from cells producing the wild-type TssJLM complex or the TssJLM complex with an unacylated variant of TssJ (Cys1-to-Ser substitution, CS) were solubilized by lauroyl sarcosine to separate inner membranes and outer membranes. $_{\text{His}}$TssM, $_{\text{Flag}}$TssL and TssJ$_{\text{Strep}}$ (indicated on the right) were revealed by anti-His, anti-Flag and anti-StrepII antibodies respectively. Controls included immuno-detection of the inner membrane DglA diacylglycrol lipase and the outer membrane OmpF porin. Wild-type TssJLM complex co-fractionates with both the inner and outer membrane fractions whereas the Cys1-to-Ser substitution mutated complex co-fractionates only with the inner-membrane fraction. Top right: negative-stain EM of the mutated TssJ$_{\text{CS}}$LM complex. Representative micrograph of the data set used for image processing. Isolated TssJ$_{\text{CS}}$LM complexes were clearly visible (white circles). Bottom right: gallery of representative class averages generated after reference-free 2D classification in Relion. End to side views are shown from top left to bottom right. **c**, Orientation of the TssJLM complex in the cell envelope. Left: schematic representation of the TssJ (J, orange), TssL (L, blue) and TssM (M, green) proteins. Their localization, main characteristics (lipidation or transmembrane segments shown in black) and the location of the 6×His and StrepII tags (red balls) are indicated. The strepII and 6×His tags were introduced at the C terminus and N terminus of TssJ and TssM respectively. Middle: immune and Nanogold labelling coupled to EM. Anti-StrepII or Nanogold-NTA were incubated with the TssJLM complex and visualized by negative-stain EM. A gallery of representative views is presented (top row, StrepII labelling; bottom row, Ni-NTA labelling). StrepII antibodies (a schematic diagram with StrepII antibodies depicted as blue circles is shown on top) and Nanogold-NTA are highlighted in red circles. Right: the positions of the StrepII antibody (targeting TssJ C terminus) and of the Ni-NTA gold particle (targeting TssM N terminus) are indicated on the TssJLM reconstruction. **d**, SAXS data and low-resolution structure of the TssM$_{\text{p}}$–TssJ complex. Top left: experimental scattering data (green crosses) and the fitting curve (continuous red line) calculated from an *ab initio* model of the TssM$_{\text{p}}$–TssJ complex. Top right: Guinier plot (dots) with the linear fit (continuous line). Bottom left: distance distribution function of the TssM$_{\text{p}}$–TssJ complex. Bottom right: SAXS envelope (grey surface) of the 'best representative' model of the TssM$_{\text{p}}$–TssJ complex. Each view is rotated by 90° around the *y*-axis. **e**, Location of the TssMp–TssJ complex SAXS envelope in the 3D reconstruction of TssJLM complex. Left: the volume of the TssM$_{\text{p}}$–TssJ complex determined by SAXS was docked into the EM 3D reconstruction of the TssJLM complex (top). Two optimal docking positions were found, both with 82% correlation with the EM map (coloured magenta and pink). The corresponding volumes in the EM map were extracted (bottom). They correspond to the same volume displayed in Extended Data Fig. 4a. Right: direct comparison of the SAXS (magenta) and EM (blue and green) volumes corresponding to the TssM$_{\text{p}}$–TssJ complex. The volumes are equivalent in size and shape. **f**, TssL cytoplasmic domain docking into the TssJLM complex base. Fitting of the TssL cytoplasmic domain (TssL$_{\text{cyto}}$)[28] dimer in green ribbons in the hooks found in the base. Top and bottom: side and bottom views, respectively.


a
MNKLACLSGRFGRPGIVFIGVAALWWLITRYGAYLGAETRRDQILLLILLSLGVLFVCYLPVMKKYVQELTYRRR
ARKEQRLPDDEERLAQTPPRYVTVQDIRHTLRRQYGRFWGRKIRILLITGTASEVELLTPGLTEQFWQEEQGTLL
LWGGDPSQPENADWLAALRRLRYRPADGIVWVTSGLSETLSAPLTEDALDRVSRAVSSCCERLGWRLPLYVWSLQ
ESPDERGRITQPVGCLLPAECSSDKLKAQLQAMLPGLVAQGIQQICCAPRYYFLLSLAERFRRNIDAVVEPLSVL
LRPYRQLLLAGIVFSPATVGGERSVRHRWRMDNRWEALPETVQQLPVRLQPSRTGHNWRRSLAVMAAILMMAQGT
GMVVSFLANRSLVAEVQEQIRPAQNQQLSPAERLQALLNLQKSLARLQYREEHGAPWYLRAGMNQNADLLAVVMP
LYAQNAHLLLRDAAAAHLEQQLRTFIRLPPDSPQRGKMAKAAYDQLRLYLMLAQPQHMEPAWFSRTLMREWPQRD
GVSAVFWQANGPTLLAYYASGIITHPQWKLTADEELVSQSRTLLLRHLGTQNSDAMLYQKMLARVAHQFADMRLT
DMTGDTDVSRLFFTDEVVPGMFTRQAWEEAVLPSIDTVINERREEMDWVLTDGRQKAPSPVSPEALRQRLTTRYF
ADFGNAWLNFLNSLHLRKAQTLSDVTEQLTLMADVRQSPLVALMNTLAVQGCTGQPREAVTDSLVKSARNLLSQE
KQPVAVPESRLHGPLATTFGPVLALMDNQNNSADMLNLQTYLTRVTQVRLRLQQIAGSSDPQAMMQLLAQTVLQG
KSVDLTDTRDYGSLTAAGLGQEWYGFGQTVFVRPMEQAWQQVLTPAAESLNARWRTAVVDGWNNAFSGRYPFKNV
SSDASLPLLAKYLNTDTGRIARFLQNNLSGVLHREGSRWVPDTINTRGLTFNPAFLKAINTLSEIADVAFTTGNA
GLHFELRPGTAAGVMQTTLITDNQKLIYVNQMPVWKRFTWPADTEAPGASLSWVSTQAGTRQYADLPGSWGLIRL
LEMARRKAAPGVASGWSLSWQAQDGRMLNYTLRTEAGEGPLVLLKLRNFVLPETVFELSGTSAFTGNDEDAGDTV
EETD
TM helices
cytoplasmic domain (64-360)
periplasmic N-terminal domain (386-835)
periplasmic C-terminal domain Nt32(836-1129)
periplasmic C-terminal domain Nt26(868-1107)
b
CDR1
CDR2
CDR3
90°
Y105
Y114
Y109
β1
β2
β3
β4
β5
β7
β8
β9
β10
β11
α1
α2
α3
α4
α5
Ct
Nt
c
Ct
Nt
Loop 1-2
90°
α1
α2
α3
α5
β1
β2
β3
β4
β5
β6
β7
β8
β9
β10
β11
Ct
Nt
TssJ
L5-6
P116
L1-2
I43
S39
N37
β1
β5
Y65
L114
L3-4
L45
β4
β2
S46
V48
M112
β3
Q89
W87
TssM
TssM
V1029
Q991
V1004
Y1003
M990
N1005
V1009
P1008
T1031
Q1006
M1007
TssJ
A986
T985
d
e
lipid
anchor
OM
24 residues
f
top view
side view

Extended Data Figure 5 | Crystal structure of the TssJ–TssM C-terminal domain complex. **a**, Amino-acid sequence of TssM. The different domains as well as the fragments used in this study are indicated (yellow, transmembrane helix; grey, cytoplasmic domain; green, blue and purple, periplasmic domain; blue and purple, C-terminal domain corresponding to the $\text{TssM}_{32\text{Ct}}$ fragment; purple, C-terminal domain corresponding to the $\text{TssM}_{26\text{Ct}}$ fragment). **b**, Crystal structure of the $\text{TssM}_{32\text{Ct}}$–nb25 complex. The two proteins are represented as rainbow-coloured ribbons. The complementary determining regions (CDRs 1–3, coloured blue, green and red, respectively) of nb25 are indicated. The inset highlights the $\text{TssM}_{32\text{Ct}}$–nb25 interface: the $\text{TssM}_{32\text{Ct}}$ surface is coloured beige whereas nb25 is represented as rainbow-coloured ribbons; the side chains of the amino acids in contact with $\text{TssM}_{32\text{Ct}}$ are indicated. The nb25 nanobody binds the TssM C-terminal domain, and covers a surface area of 580 Å^2 by inserting its protruding CDR3 between $\text{TssM}_{32\text{Ct}}$ loops L5–6 and L9–10. The contacts between the two proteins are listed in Extended Data Table 2a. **c**, Crystal structure of the $\text{TssM}_{26\text{Ct}}$–TssJ complex. Left: the two proteins are represented as ribbons and coloured in rainbow mode. Middle: same view rotated by 90°. The TssJ loop 1–2, previously shown to contact TssM[25], is indicated. Right: $\text{TssM}_{26\text{Ct}}$–TssJ interface. Top panel: the $\text{TssM}_{26\text{Ct}}$ surface is coloured violet, whereas TssJ is represented as rainbow-coloured ribbons. The TssJ side-chains of the amino acids in contact with TssM are indicated. The loops are numbered according to the flanking β-strands. Bottom panel: the TssJ surface is coloured beige whereas $\text{TssM}_{26\text{Ct}}$ is represented as rainbow-coloured ribbons. The TssM side-chains of the amino acids in contact with TssJ are indicated. The contacts between the two proteins are listed in Extended Data Table 2b. **d**, Comparison of the binding sites of nb25 and TssJ on TssM. Left: the structure of the $\text{TssM}_{26\text{Ct}}$–TssJ complex (rainbow coloured) has been superimposed to the structure of the $\text{TssM}_{32\text{Ct}}$–nb25 complex (only nb25 is shown in grey for clarity). Right: the same partners as in the left panel in surface representation. $\text{TssM}_{26\text{Ct}}$ (violet), TssJ (green) and nb25 (pink). **e**, Insertion of the TssJ lipid anchor in the outer membrane. Left: TssJ structure[25] with the N-terminal 24 residues (absent in the crystal structure). This N-terminal extension (in magenta), predicted to be disordered, was modelled in Chimera using Modeller. The first cysteine residue is acylated to allow anchorage to the inner leaflet of the outer membrane (orange rectangle). Right: docking of the $\text{TssM}_{26\text{Ct}}$–TssJ complex in the EM 3D reconstruction of the TssJLM complex (only the uppermost (tip) part of the TssJLM complex is shown). Left panel: two $\text{TssM}_{26\text{Ct}}$–TssJ were docked into the inner and outer pillars of the tip complex. Right panel: docking in each pillar of the TssJLM tip complex (C_5 symmetry). **f**, Hydrophobicity of the $\text{TssM}_{26\text{Ct}}$–TssJ complex. Surface representation of the $\text{TssM}_{26\text{Ct}}$–TssJ decamer (left, top view; right, side view). The hydrophobicity of the surface residues is displayed (blue to red scale from most hydrophilic to most hydrophobic). No obvious hydrophobic patch is visible at the surface of the complex.


a
WT
ΔtssM
MWT
MT972C
MV989C
MN1005C
MT1019C
MT1035C
MV1062C
MG1075C
MV1092C
MS1109C
C S C S C S C S C S C S C S C S C S C S C S C S
50
36
30
16
TolB
Hcp
b
WT
ΔtssBC
M
WC
BSA-mal
+ - + - +
250
98
T972C
V989C
N1005C
T1019C
S1109C
T1035C
V1062C
G1075C
V1092C
TssM
*
ΔtssJ
WC
- +
c
1005
989
90°

Extended Data Figure 6 | Cell surface accessibility of TssM C-terminal domain. **a**, Functionality of the TssM cysteine variants. Hcp release was assessed by separating whole cells (C) and supernatant (S) fractions from the wild-type (WT) 17-2 strain and its Δ*tssM* derivative producing a wild-type allele of TssM or TssM cysteine substitution derivatives (as indicated). A total of 1×10^9 cells and the TCA-precipitated material from the supernatant of 2×10^9 cells were analysed by western blot using anti-HA monoclonal antibody (lower panel) and anti-TolB polyclonal antibodies as a lysis control (upper panel). The molecular mass markers (in kilodaltons) are indicated on the left. **b**, Cysteine substitution labelling. Accessibility to cysteine residues positioned in TssM domain 4 loops was assessed by treating isolated membranes (M) or whole cells (WC) of the indicated strain (WT, wild-type 17-2; Δ*tssBC*; Δ*tssJ*) producing the indicated TssM cysteine derivative (in red letters) with the cysteine-reactive, membrane-impermeant BSA-maleimide (BSA-mal). Samples corresponding to a total of 5×10^9 cells were analysed by western blot using anti-Flag monoclonal antibody. The position of the TssM protein (~125 kDa) is indicated as well as that of a retarded band corresponding to BSA-maleimide-coupled TssM (~190 kDa; asterisk). The molecular mass markers (in kilodaltons) are indicated. **c**, Close-up of the $\text{TssM}_{26\text{Ct}}$–TssJ interface. $\text{TssM}_{26\text{Ct}}$ is represented in blue ribbons. TssJ is represented in orange ribbons and orange transparent surface. TssM residues accessible from the cell exterior when the T6SS is functional are indicated by yellow spheres whereas unaccessible residues are shown by grey spheres. The accessible residues 989 and 1005 are buried at the interface between TssM and TssJ, suggesting that this interface is probably disrupted during T6SS assembly and/or function. Left and right panels are orthogonal views of the same molecule.



Extended Data Figure 7 | Comparison with other bacterial secretion systems and model for channel opening. **a**, Comparison between the T6SS TssJLM membrane core complex structure and other bacterial secretion systems. From left to right, the *E. coli* AcrAB-TolC multi-drug efflux pump (EMDB accession number emd-5915)[73], the EAEC T6SS membrane core complex (this study, EMDB accession number emd-2927), the *Shigella* T3SS transmembrane complex (EMDB accession number emd-1617)[74] and the *E. coli* R388 T4SS complex (EMDB accession number emd-2567)[75]. The positions on the inner membrane (IM) and outer membrane (OM) are indicated (C, cytoplasm; P, periplasm). Scale bar, 10 nm. **b**, Docking of the Hcp tube/VgrG spike into the TssJLM 3D reconstruction. Left: before sheath contraction. The Hcp tube/VgrG spike (VgrG in yellow and Hcp in green; surface representation) was manually docked in the 3D reconstruction of TssJLM complex (grey surface). The diameter of the channel defined by the closed tip complex is not large enough to allow the passage of the tube/spike, suggesting that large conformational changes probably occur. The cavity at the tip of VgrG could be filled by VgrG-bound PAAR modules or toxin effectors[20]. Right: during sheath contraction. The diameter of the C_{10}-symmetrized $TssM_{26Ct}$ model (represented as ribbons) is compatible with the passage of the Hcp tube/VgrG spike (same colours as in the left panel). **c**, Closed and open forms of the $TssM_{26Ct}$ oligomer. Crystal structure of $TssM_{26Ct}$ represented as ribbons and transparent surface. The $TssM_{26Ct}$ α- and β-domains are coloured cyan and blue, respectively. The C-terminal α5-helix and the extended stretch are coloured pink. Cysteines with extracellular accessibility when the T6SS is active are coloured yellow, while the unlabelled ones are coloured red. Left: docking of the $TssM_{26Ct}$–TssJ crystal structure in the EM 3D reconstruction of the TssJLM tip complex. Top and bottom panels, side and top views, respectively. Right: model of a C_{10}-symmetrized oligomer of the $TssM_{26Ct}$ domain. Top and bottom panels, side and top views, respectively.

Extended Data Table 1 | Data collection and refinement statistics

	TssM_{32Ct}-nb25	TssM_{26Ct}	TssM_{26Ct}-TssJ
Data collection§			
Space group	$P6_4$	$P\,4_12_12$	$P\,4_12_12$
Cell dimensions			
a, *b*, *c* (Å)	95.2, 95.2, 172.95	64.0, 64.0, 249.7	85.5, 85.5, 256.4
α, *β*, *γ* (°)	90.0, 90.0, 120.0	90.0, 90.0, 90	90.0, 90.0, 90
Resolution (Å)	50.0-1.92(1.97-1.92)*	30.0-1.51(1.6-1.51)*	50.0-2.24(2.38-2.24)*
R_{merge}	0.079 (1.08)	0.067 (0.59)	0.067 (0.73)
$I/\sigma I$	18.0 (2.0)	19.2 (3.0)	21.5 (3.1)
Completeness (%)	100.0 (100.0)	99.9 (89.3)	99.7 (98.4)
Redundancy	11.4 (11.3)	9.9 (9.9)	10 (10)
Refinement			
Resolution (Å)	47.6-1.92(1.97-1.92)*	22.1-1.51 (1.55-1.51)*	49.3-2.24(2.3-2.24)*
No. reflections	67543 (4721)	82127 (5359)	46047 (3015)
$R_{\text{work}}/R_{\text{free}}$	0.184/0.21(0.234/0.25)	0.192/0.202(0.241/27.6)	0.208/0.228(0.224/0.25)
No. atoms			
Protein	5522	3784	5521
Ligand/ion	15	22	4
Water	805	536	379
B-factors			
Protein	42.3	27.4	51.7
Ligand/ion	98.8	55.0	81
Water	49.9	39.5	63.8
R.m.s deviations			
Bond lengths (Å)	0.009	0.010	0.010
Bond angles (º)	1.05	1.03	1.19

§Each data set has been collected on a unique crystal.
*Highest resolution shell is shown in parenthesis.

Extended Data Table 2 | Interactions and accessibility data

Extended Table 2a. Interactions between TssM and nb25

n25/CDR3			TssM$_{26Ct}$				
number	type	atom	number	type	atom	distance (Å)	bond
103	Gly	O	1063	Ala	N	2.89	H
104	Ile	CA	1061	Gly	O	3.08	
105	Tyr	N	1061	Gly	O	2.85	H
107	Thr	OG1	1060	Pro	O	2.67	H
107	Thr	CG2	1061	Gly	CA	3.56	
109	Tyr	CE1	1067	Ser	CB	3.50	
109	Tyr	OH	1080	Tyr	O	3.47	H
109	Tyr	OH	1081	Thr	OG1	2.64	H
110	Ile	CD1	1062	Val	CG2	3.73	
113	Pro	O	984	Gly	CA	3.12	
113	Pro	O	985	Thr	N	3.00	H
114	Tyr	CE1	982	Arg	CD	3.55	
114	Tyr	CZ	982	Arg	NE	3.63	
114	Tyr	OH	1010	Trp	NE1	2.99	H
114	Tyr	O	982	Arg	NE	2.88	H
115	Gly	O	982	Arg	NH2	2.76	H
116	Met	O	1008	Pro	CD	3.32	
117	Asp	OD1	982	Arg	NH2	2.80	H

Extended Table 2c. TssM cysteine accessibility.

Cysteine position	Labelled at rest	Labelled in action	Position	WAS #
972	-	-	helix α3	107 / 107
989	-	++	loop β3- β4	14 / 10
1005	-	++	loop β5- β6	76 / 0
1019	-	-	loop β6- β7	35 / 35
1035	-	+	loop β7- β8	113 / 113
1062	-	-	loop β11- α5	94 / 94
1075	-	++	loop β10- β11	51 / 51
1092	-	-	helix α5	0 / 0
1109	++	++	C-terminus	NA*

Extended Table 2b Interactions between TssM and TssJ.

TssJ			TssM			distance(Å)
37	Asn	Nd2	1005	Asn	O	3.06 H
39	Ser	Cb	985	Thr	Og1	3.80
43	Ile	Cg2	987	Ala	Cb	3.79
45	Leu	Cd1	985	Thr	Cg2	3.63
		Cb	1005	Asn	Nd2	3.56
46	Ser	Og	990	Met	Ca	3.14
		O	1005	Asn	Nd2	3.27 H
		Og	1005	Asn	Nd2	3.19 H
48	Val	Cg2	1004	Val	Cb	3.82
65	Tyr	Ce1	1007	Met	Cg	3.51
		Oh	1007	Met	N	3.30 H
87	Trp	Ch2	990	Met	C	3.56
89	Gln	Oe1	1031	Thr	O	2.87 H
		Ne2	1031	Thr	Og1	2.79 H
112	Met	Ce	1006	Gln	Ne2	3.35
113	Phe	O	1007	Met	N	3.24 H
114	Leu	Cd1	1003	Tyr	OH	3.16
		N	1005	Asn	O	3.33 H
		Cd1	1006	Gln	O	3.40
		O	1007	Met	Cg	3.51
		Cd1	1008	Pro	Cd	3.73
116	Pro	Cd	1007	Met	Cg	3.74

TssJ

Loop L1,2

Loop L3,4

Loop L5,6

TssM

Loop L3,4

Loop L5,6

The letter H in right-hand columns indicates that atoms establish a hydrogen bond.
WAS, water-accessible surface of the original amino acids (measured in the unbound TssM/in the TssM–TssJ complex).
*Not visible in the electron density map.

ARTICLE

doi:10.1038/nature14656

Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser

Yanyong Kang[1]*, X. Edward Zhou[1]*, Xiang Gao[1]*, Yuanzheng He[1]*, Wei Liu[2], Andrii Ishchenko[3], Anton Barty[4], Thomas A. White[4], Oleksandr Yefanov[4], Gye Won Han[3], Qingping Xu[5], Parker W. de Waal[1], Jiyuan Ke[1], M. H. Eileen Tan[1,6], Chenghai Zhang[1], Arne Moeller[7], Graham M. West[8], Bruce D. Pascal[8], Ned Van Eps[9]†, Lydia N. Caro[10], Sergey A. Vishnivetskiy[11], Regina J. Lee[11], Kelly M. Suino-Powell[1], Xin Gu[1], Kuntal Pal[1], Jinming Ma[1], Xiaoyong Zhi[1], Sébastien Boutet[12], Garth J. Williams[12], Marc Messerschmidt[12,13], Cornelius Gati[4], Nadia A. Zatsepin[2,14], Dingjie Wang[2,14], Daniel James[2,14], Shibom Basu[2,14], Shatabdi Roy-Chowdhury[2,14], Chelsie E. Conrad[2], Jesse Coe[2], Haiguang Liu[2,15], Stella Lisova[2], Christopher Kupitz[2,16], Ingo Grotjohann[2], Raimund Fromme[2], Yi Jiang[17], Minjia Tan[17], Huaiyu Yang[17], Jun Li[6], Meitian Wang[18], Zhong Zheng[19], Dianfan Li[20], Nicole Howe[20], Yingming Zhao[13,21], Jörg Standfuss[22], Kay Diederichs[23], Yuhui Dong[24], Clinton S. Potter[7], Bridget Carragher[7], Martin Caffrey[20], Hualiang Jiang[17], Henry N. Chapman[4,25], John C. H. Spence[2,14], Petra Fromme[2], Uwe Weierstall[2,14], Oliver P. Ernst[10,26], Vsevolod Katritch[19], Vsevolod V. Gurevich[11], Patrick R. Griffin[8], Wayne L. Hubbell[9], Raymond C. Stevens[3,19,27], Vadim Cherezov[3], Karsten Melcher[1] & H. Eric Xu[1,28]

G-protein-coupled receptors (GPCRs) signal primarily through G proteins or arrestins. Arrestin binding to GPCRs blocks G protein interaction and redirects signalling to numerous G-protein-independent pathways. Here we report the crystal structure of a constitutively active form of human rhodopsin bound to a pre-activated form of the mouse visual arrestin, determined by serial femtosecond X-ray laser crystallography. Together with extensive biochemical and mutagenesis data, the structure reveals an overall architecture of the rhodopsin–arrestin assembly in which rhodopsin uses distinct structural elements, including transmembrane helix 7 and helix 8, to recruit arrestin. Correspondingly, arrestin adopts the pre-activated conformation, with a ~20° rotation between the amino and carboxy domains, which opens up a cleft in arrestin to accommodate a short helix formed by the second intracellular loop of rhodopsin. This structure provides a basis for understanding GPCR-mediated arrestin-biased signalling and demonstrates the power of X-ray lasers for advancing the frontiers of structural biology.

G-protein-coupled receptors (GPCRs) comprise the largest family of cell surface receptors, which signal primarily via G proteins or arrestins[1,2]. Upon activation, GPCRs recruit heterotrimeric G proteins and subsequently G-protein-coupled receptor kinases (GRKs), which phosphorylate GPCRs to allow the high-affinity binding to arrestin[3]. Arrestin binding to the receptors blocks their interactions with G proteins and leads to the receptor's desensitization[4]. The binding of arrestins to GPCRs also initiates numerous cellular signalling pathways that are independent of G proteins. Arrestin-mediated signalling is therefore a central component of the GPCR functional network.

GPCRs are targets of one-third of the current clinically used drugs. Recent studies have demonstrated that G-protein and arrestin pathways are distinct and can be pharmacologically modulated independently using biased GPCR ligands[5]. Biased GPCR ligands are often preferred over unbiased agonists and antagonists, as they selectively direct the receptor to a subset of partners and can deliver therapeutic benefits with fewer undesirable side effects. Research towards biased ligands has become a new trend for GPCR-targeting therapeutics[6].

The molecular mechanisms of GPCR signalling have been unravelled by recent breakthroughs in GPCR structural biology[7–10]. In the antagonist-bound state, GPCRs assume a closed conformation with the cytoplasmic ends of the transmembrane (TM) helices packed closely with each other[7,9], thus blocking the interactions with G proteins or arrestins. In contrast, agonist binding promotes conformational changes in GPCRs, including a dramatic movement within the cytoplasmic side

[1]Laboratory of Structural Sciences, Center for Structural Biology and Drug Discovery, Van Andel Research Institute, Grand Rapids, Michigan 49503, USA. [2]Department of Chemistry and Biochemistry, and Center for Applied Structural Discovery, Biodesign Institute, Arizona State University, Tempe, Arizona 85287-1604, USA. [3]Department of Chemistry, Bridge Institute, University of Southern California, Los Angeles, California 90089, USA. [4]Center for Free Electron Laser Science, Deutsches Elektronen-Synchrotron DESY, 22607 Hamburg, Germany. [5]Joint Center for Structural Genomics, Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA. [6]Department of Obstetrics & Gynecology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore. [7]The National Resource for Automated Molecular Microscopy, New York Structural Biology Center, New York, New York 10027, USA. [8]Department of Molecular Therapeutics, The Scripps Research Institute, Scripps Florida, Jupiter, Florida 33458, USA. [9]Jules Stein Eye Institute and Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, USA. [10]Department of Biochemistry, University of Toronto, Toronto, Ontario M5S 1A8, Canada. [11]Department of Pharmacology, Vanderbilt University, Nashville, Tennessee 37232, USA. [12]Linac Coherent Light Source (LCLS), SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA. [13]BioXFEL, NSF Science and Technology Center, 700 Ellicott Street, Buffalo, New York 14203, USA. [14]Department of Physics, Arizona State University, Tempe, Arizona 85287, USA. [15]Beijing Computational Science Research Center, Haidian District, Beijing 10084, China. [16]Department of Physics, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53211, USA. [17]State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China. [18]Swiss Light Source at Paul Scherrer Institute, CH-5232 Villigen, Switzerland. [19]Department of Biological Sciences, Bridge Institute, University of Southern California, Los Angeles, California 90089, USA. [20]School of Medicine and School of Biochemistry and Immunology, Trinity College, Dublin 2, Ireland. [21]Ben May Department for Cancer Research, University of Chicago, Chicago, Illinois 60637, USA. [22]Laboratory of Biomolecular Research at Paul Scherrer Institute, CH-5232 Villigen, Switzerland. [23]Department of Biology, Universität Konstanz, 78457 Konstanz, Germany. [24]Beijing Synchrotron Radiation Facility, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China. [25]Centre for Ultrafast Imaging, 22761 Hamburg, Germany. [26]Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada. [27]iHuman Institute, ShanghaiTech University, 2F Building 6, 99 Haike Road, Pudong New District, Shanghai 201210, China. [28]VARI-SIMM Center, Center for Structure and Function of Drug Targets, CAS-Key Laboratory of Receptor Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China. †Present address: Department of Biochemistry, University of Toronto, Toronto, Ontario M5S 1A8, Canada.
*These authors contributed equally to this work.

of the TM domain[8,11–14], thus allowing activated receptors to recruit G proteins or arrestins to mediate downstream signalling. However, arrestin coupling to GPCRs may require a conformation of the receptor different from that required for coupling with G proteins[14,15].

Rhodopsin is a prototypical GPCR responsible for light perception[7]. Along with the β_2-adrenergic receptor (β_2AR), rhodopsin has served as a model system for studying GPCR signalling[16]. Figure 1a shows rhodopsin binding to G protein and arrestin. Light induces isomerization of 11-*cis* retinal to all-*trans*-retinal (ATR), which activates rhodopsin and promotes its interactions with G protein[17,18]. Light-activated rhodopsin is then phosphorylated by rhodopsin kinase (GRK1), leading to high affinity recruitment of arrestin that terminates the G protein signalling. Activation of rhodopsin can also be achieved through mutations, including the E113$^{3.28}$Q/M257$^{6.40}$Y mutation, which yields a constitutively active rhodopsin[19] (superscripts in residues refer to the Ballesteros–Weinstein numbering[20]). The crystal structure of bovine rhodopsin has been determined in the inactive, resting state[7], the ligand-free state[21,22], and the ligand-activated state in complex with a G-protein peptide[23]. Arrestin structures have been determined in the inactive[24,25] and pre-activated form[1,2]. Recent electron microscopy analysis has revealed the assembly and conformational dynamics of the β_2AR–β-arrestin complex[26]. Here we report the crystal structure of an active form of human rhodopsin bound to a pre-activated mouse visual arrestin, determined by serial femtosecond crystallography (SFX). The structure has been confirmed by electron microscopy, double electron-electron resonance (DEER) spectroscopy, hydrogen–deuterium exchange mass spectrometry (HDX), cell-based rhodopsin–arrestin interaction assays, and site-specific disulfide crosslinking experiments. Our study provides a molecular basis for understanding GPCR-mediated arrestin-biased signalling.

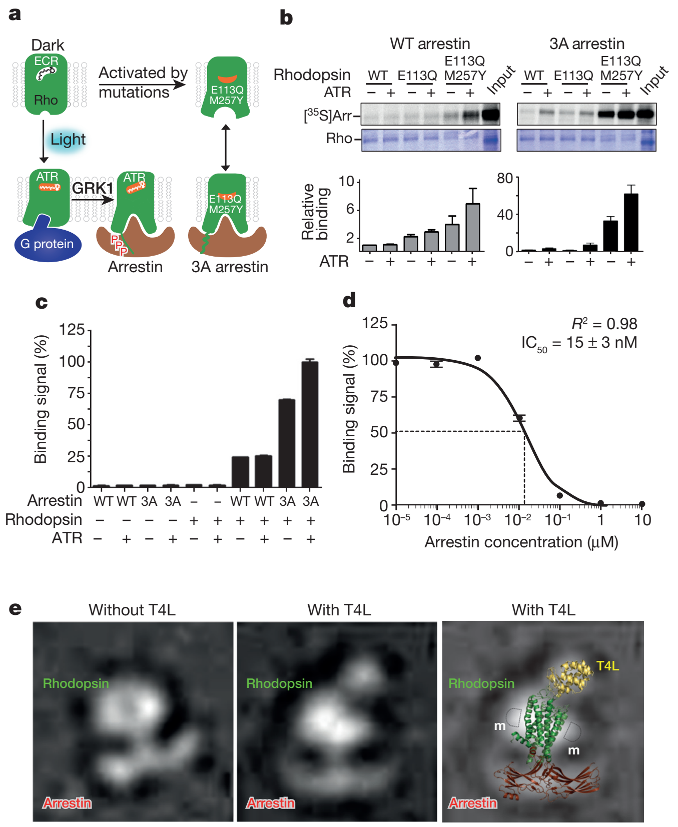


Figure 1 | Rhodopsin–arrestin interactions and complex assembly. **a**, Diagram of the binding of rhodopsin (Rho) with G protein and arrestin as described in the main text. Labels are 11-*cis* retinal (ECR), all-*trans*-retinal (ATR), and rhodopsin kinase (GRK1). **b**, Rhodopsin (Rho) and arrestin (Arr) interaction determined by pull-down assay in the absence and presence of ATR (top panel). Middle panel, rhodopsin loading controls. Bottom panel, relative binding of ^{35}S-labelled arrestin was determined by densitometry ($n = 3$, error bars, s.d.). **c**, Binding of His$_8$–MBP–rhodopsin (E113$^{3.28}$Q/M257$^{6.40}$Y) protein to biotin–MBP–arrestin (wild type (WT) and 3A) measured by AlphaScreen in the absence or presence of 5 μM ATR. The first six columns are controls (luminescence signals in the presence of only one of the binding partners; $n = 3$, error bars, s.d.). **d**, Competition of arrestin binding to rhodopsin was determined by a homologous AlphaScreen assay and the IC$_{50}$ value was derived from repeat experiments ($n = 3$, error bars, s.d.). **e**, Negative stain electron microscopy images of rhodopsin–arrestin complexes without or with T4L at the N terminus; right panel, overlay of the electron microscopy image with the structures of T4L, rhodopsin and arrestin. m, detergent micelle.

Characterization and crystallization

To characterize the rhodopsin–arrestin interaction, we expressed and purified E113$^{3.28}$Q and E113$^{3.28}$Q/ M257$^{6.40}$Y mutant receptors (Extended Data Fig. 1a). These mutations were introduced in the context of the N2$^{\mathrm{Nterm}}$C/N282$^{\mathrm{ECL3}}$C mutant that is known to create a disulfide bond that increases rhodopsin stability without affecting its activity[27–29]. The N2$^{\mathrm{Nterm}}$C/N282$^{\mathrm{ECL3}}$C mutant is referred to as our wild-type control. To determine the interaction between rhodopsin and arrestin, we developed a bead binding pull-down assay. In this assay, rhodopsin expressed as a fusion with a maltose-binding protein (MBP) at its N terminus was bound to amylose beads, which were then used to pull down *in vitro* translated arrestin labelled with ^{35}S. Wild-type arrestin has weak background binding to wild-type rhodopsin (Fig. 1b). The E113$^{3.28}$Q mutation increased wild-type arrestin binding by twofold to threefold, and the E113$^{3.28}$Q/ M257$^{6.40}$Y mutation further increased the binding of wild-type arrestin in the presence of all-*trans*-retinal (fourfold to eightfold). In contrast to the relatively weak binding of wild-type arrestin, the binding of 3A arrestin, a pre-activated form of arrestin that obviates the need for receptor phosphorylation for high affinity binding through three alanine mutations in L374, V375, and F376 in the C-terminal tail of arrestin, is much stronger. In the absence of all-*trans*-retinal, we observed a nearly 30-fold increase of 3A arrestin binding to the E113$^{3.28}$Q/M257$^{6.40}$Y receptor. All-*trans*-retinal further increased 3A arrestin binding to the E113$^{3.28}$Q/M257$^{6.40}$Y receptor by ~60-fold above the binding of wild-type arrestin to wild-type rhodopsin (Fig. 1b and Extended Data Fig. 1b).

We also measured rhodopsin–arrestin interactions using AlphaScreen assays (Extended Data Fig. 1c) with His$_8$-tagged rhodopsin and biotin-tagged arrestin. Wild-type arrestin interacted weakly with the E113$^{3.28}$Q/M257$^{6.40}$Y rhodopsin, regardless of the presence of all-*trans*-retinal (Fig. 1c). As a positive control, the GαCT-HA peptide, a high affinity peptide variant of the C terminus of G-transducin (G$_t$)[30], readily interacted with the E113$^{3.28}$ Q/M257$^{6.40}$Y receptor in the absence of all-*trans*-retinal, and addition of all-*trans*-retinal slightly increased this interaction (Extended Data Fig. 1d). Quantitative competition using unlabelled 3A arrestin or GαCT-HA with the E113$^{3.28}$Q/M257$^{6.40}$Y receptors revealed an IC$_{50}$ value of 15 nM and 700 nM for the binding of 3A arrestin and the GαCT-HA peptide, respectively (Fig. 1d and Extended Data Fig. 1d). The strength of the interaction between the E113$^{3.28}$Q/M257$^{6.40}$Y rhodopsin and the 3A arrestin is in a similar range as the estimated K_d value of 30–80 nM for the binding of arrestin to the fully activated phosphorylated rhodopsin[31].

Mixing individually purified proteins did not yield a stable 1:1 complex, nor did it lead to crystallization. Extensive biochemical data support a 1:1 stoichiometry in the rhodopsin–arrestin complex[32,33]. Therefore, we engineered a fusion protein in which 3A arrestin is linked by a 15-residue linker to the C terminus of E113$^{3.28}$ Q/M257$^{6.40}$Y rhodopsin. We expressed and purified the rhodopsin–arrestin fusion protein, as well as a T4 lysozyme (T4L)–rhodopsin–arrestin fusion, in which a T4L is fused to the N terminus of rhodopsin to increase the soluble surface for crystallization (Extended

Data Fig. 2a). The T4L–rhodopsin–arrestin fusion protein is monomeric and relatively stable with a T_{m} of 59 °C (Extended Data Fig. 2b, c). Negative stain electron microscopy images revealed that E113$^{3.28}$Q/M257$^{6.40}$Y rhodopsin and 3A arrestin form a stable complex with arrestin bound to the cytoplasmic side of rhodopsin (Fig. 1e). The T4L–rhodopsin–arrestin fusion protein formed crystals with sizes in the range of 5 to 15 μm under various lipid cubic phase (LCP) crystallization conditions (Extended Data Fig. 2d, e). Despite extensive optimization, the crystals diffracted only to 6–8 Å at synchrotron sources (Extended Data Fig. 2f). We thus turned our attention to the emerging method of SFX[34] with an LCP injector (LCP-SFX)[35,36].

Structure determination by SFX

Because of the small size of crystals, we hypothesized that diffraction by X-ray free electron laser (XFEL) at the Linac Coherent Light Source (LCLS) would improve data quality given the advantages of intense and very short XFEL pulses for micrometre-size crystals. In the LCP-SFX method, a stream of gel-like LCP with fully hydrated microcrystals runs continuously in vacuum across the 1.5-μm-diameter XFEL beam, which delivers 120 X-ray pulses per second with less than 50 fs pulse duration and sufficient intensity to capture crystal diffraction patterns with a single pulse. Within ~12 h of run time, we collected over 5 million detector frames, of which 22,262 had more than 40 diffraction spots as determined by the Cheetah hit-finding software[37]. Diffraction patterns from 18,874 crystals could be indexed and integrated using CrystFEL[38]. The data were processed according to the apparent tetragonal lattice with a large unit cell ($a = b = 109.2$ Å and $c = 452.6$ Å). The diffraction was anisotropic with resolution limits of 3.8 Å and 3.3 Å along the a^*/b^* and c^* axes, respectively (Supplementary Table 1).

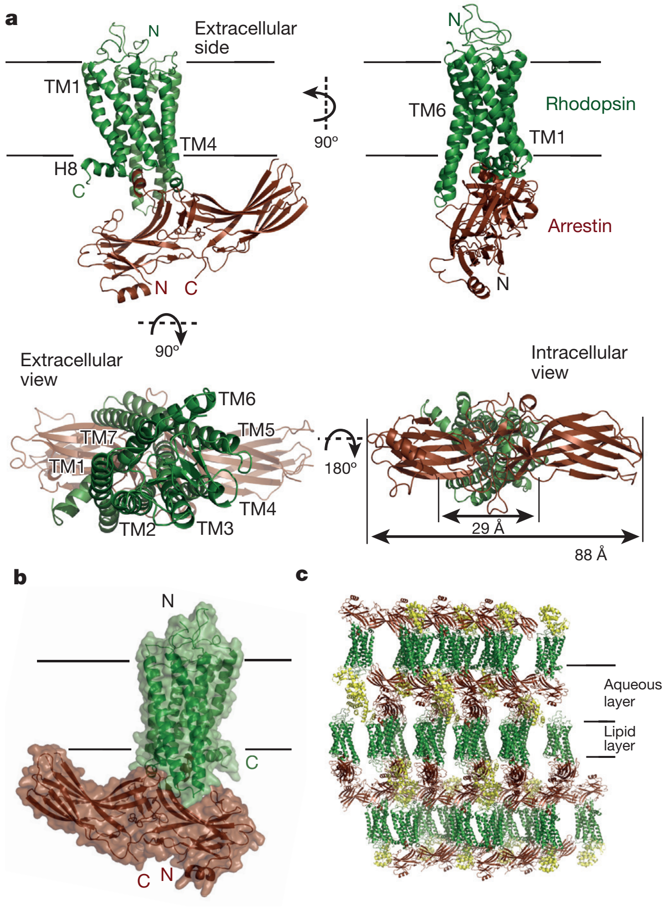


Figure 2 | The structure of the rhodopsin–arrestin complex. a, The structure of the rhodopsin–arrestin complex in four orientations. The relative dimensions of rhodopsin and arrestin are shown in the intracellular view. TM1–TM7 indicates rhodopsin transmembrane helices 1–7; H8 is intracellular helix 8. **b**, An overall view of the rhodopsin–arrestin complex shown with transparent solid surface. T4 lysozyme (T4L) is omitted from this view. **c**, Crystal packing diagram of the rhodopsin–arrestin complex with T4L as yellow ribbon model.

The crystals appeared to be pseudo-merohedrally twinned in $P2_12_12_1$ (Supplementary Table 2) and the structure was solved by molecular replacement using known structures of active rhodopsin[39] and pre-activated arrestin[1] (details in Methods). The structure contains four rhodopsins (residues 1–326), four arrestins (residues 12–361 with a small missing loop of residues 340–342), and three T4Ls (residues 2–161 in complexes A and D; residues 2–12 and 58–162 in complex C; no T4L was modelled in complex B owing to poor density) (Fig. 2). The final structure was refined to R_{work} and R_{free} of 25.2% and 29.3%, respectively, with excellent geometry (Supplementary Table 1b). The overall arrangement of the T4L–rhodopsin–arrestin complex is well supported by the electron density maps (Extended Data Fig. 3), including a 3,000 K simulated annealing omit map. Because of the twinned nature of the data sets, we performed extensive structure-validation experiments, including DEER, HDX, cell-based rhodopsin–arrestin interaction assays and site-specific disulfide cross-linking. Below we describe the rhodopsin–arrestin structure and the results of validation experiments.

Overall structure of the rhodopsin–arrestin complex

The most striking feature of the rhodopsin–arrestin complex is the asymmetric binding of arrestin to rhodopsin (Fig. 2) and this asymmetric arrangement is similar in all four complexes in the asymmetric unit, providing an independent confirmation of the rhodopsin–arrestin complex assembly (Extended Data Fig. 4). Figure 2a shows one rhodopsin–arrestin complex in four 90° orientations. From the intracellular (IC) view, rhodopsin and arrestin have similar heights, but the width of arrestin is nearly three times that of rhodopsin. Figure 2b shows the rhodopsin–arrestin complex in a transparent surface, whose overall arrangement of the domains can be fit into the electron microscopy images (Fig. 1e). Figure 2c shows the layered or type I packing of the complex in the crystal lattice with alternating hydrophilic and hydrophobic layers comprising arrestin, T4L and rhodopsin, respectively (Fig. 2c). This arrangement allows the complex to form extensive packing interactions that involve all soluble portions of the proteins, with the arrestin being the central mediator for packing with T4L, rhodopsin and arrestin from neighbouring symmetry-related molecules.

To validate the assembly of the rhodopsin–arrestin complex, we used DEER to determine intermolecular distances within the complex[40]. The DEER distances from residue Y74$^{2.41}$ of rhodopsin to three arrestin residues (T61, V140, and S241) measured in a non-fused rhodopsin–arrestin complex were 28 Å, 23 Å and 33 Å, closely matching the distances of 28 Å, 22 Å and 34 Å, respectively, as observed in the crystal structure (Fig. 3). The intramolecular distances in the active arrestin bound to light-activated phosphorylated rhodopsin have also been studied extensively by DEER[41], and all of them match exceedingly well with the crystal structure (Supplementary Table 3). Together, these data support the conclusion that the complex formed by fusion proteins closely resembles the physiologically relevant complex formed by individual proteins.

The rhodopsin–arrestin interface

The four rhodopsin–arrestin complexes in the asymmetric unit adopt nearly identical interfaces (Extended Data Fig. 4a), which are stabilized by intermolecular interactions as summarized in Supplementary Table 4. The total surface area buried in the interface is 1,350 Å^2, which is substantially smaller than the area (2,576 Å^2) buried in the β_2AR–G$_{\mathrm{s}}$ complex[8]. Unlike the continuous interface observed in the β_2AR–G$_{\mathrm{s}}$ complex, the rhodopsin–arrestin complex has four distinct

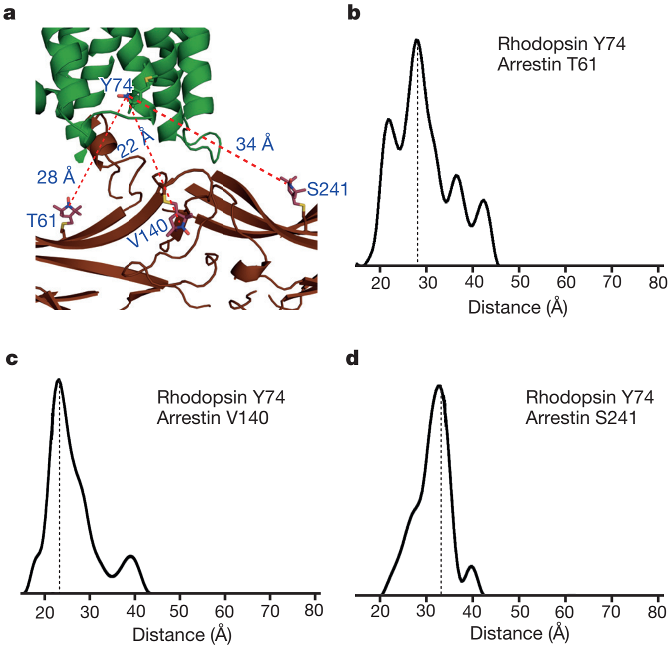


Figure 3 | DEER validation of rhodopsin–arrestin complex assembly. **a**, An overall view of rhodopsin–arrestin assembly showing the three intermolecular distances based on the models of the R1 nitroxide pairs at rhodopsin residue Y74$^{2.41}$ and three arrestin residues T61, V140, and S241 based on the crystal structure. **b**–**d**, The experimental distance distributions between the nitroxide spin labelled R1 pairs of rhodopsin Y74$^{2.41}$ and bovine arrestin S60, V139, and L240, which are in equivalent positions to mouse arrestin T61(**b**), V140 (**c**), and S241(**d**) as labelled in the figure.

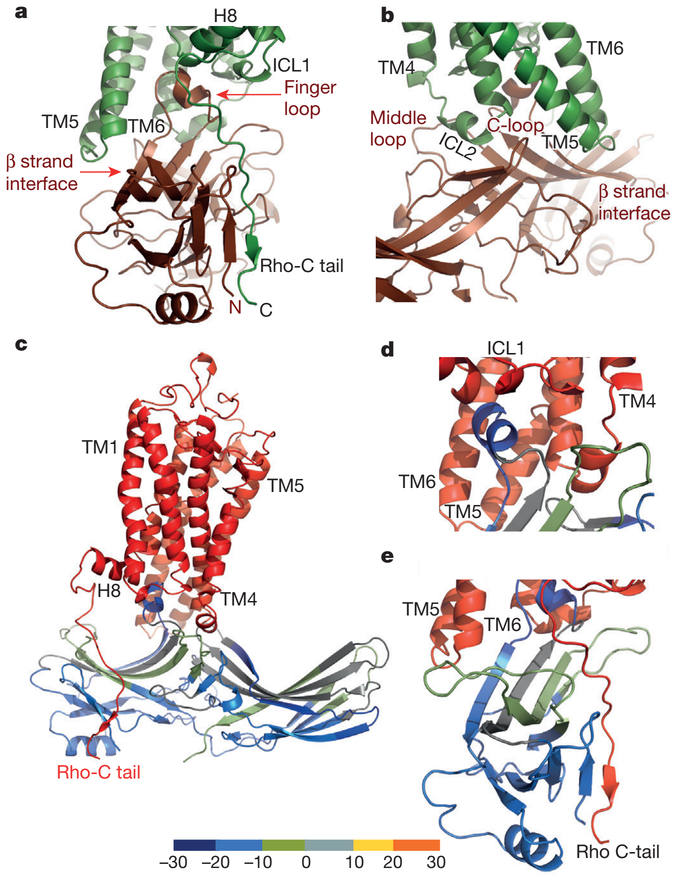


Figure 4 | The rhodopsin–arrestin interface and its validation by HDX. **a**, **b**, Two overall views showing the four interface patches of the rhodopsin–arrestin complex. **c**–**e**, Mapping of HDX on the rhodopsin-bound arrestin structure. Rhodopsin is coloured in red and arrestin is coloured based on the exchange rate differences between free 3A arrestin and rhodopsin-bound arrestin as shown in Extended Data Fig. 6a. This figure was made using a computational model of the full rhodopsin–arrestin complex.

arrestin interface patches (Fig. 4a, b and Extended Data Fig. 4b). The first arrestin interface patch consists of the finger loop (residues Q70 to L78), which adopts a short α-helix and forms extensive interactions with the C terminus of TM7 and the N terminus of helix 8, as well as the loop residues (ICL1) of rhodopsin (Fig. 5a). Interactions of arrestin with TM7 of rhodopsin are of particular interest because conformational changes in TM7 have been implicated in arrestin-biased signalling[14,15]. Moreover, the close interactions between rhodopsin's helix 8 and arrestin have been shown to be essential for high-affinity binding of arrestin to the activated rhodopsin[42]. The second arrestin interface patch is formed by the middle loop (residue V140 region) and the C-loop (residue Y251 region at the central loop in the arrestin C domain) that interact with the ICL2 of rhodopsin, and the arrestin back loop (R319 and T320) that interacts with the C terminus of TM5. The middle and C-loops are close to each other in the inactive arrestin, but move apart upon its activation to form a cleft that accommodates the ICL2 of rhodopsin, which adopts a short helix (Fig. 4a, b). The positions of the finger loop and the C-loop are supported by a composite omit $2F_o - F_c$ electron density map (Extended Data Fig. 3a, d). The third arrestin interface patch is the β-strand (residues 79–86), which follows the finger loop and interacts with residues from TM5, TM6 and ICL3 of rhodopsin. The fourth arrestin putative interface patch is mostly between its N-terminal β-strand (residues 11–19) and the C-terminal tail of rhodopsin, which was not visible in the electron density map owing to the apparent flexibility of this region, but was computationally modelled based on HDX and disulfide cross-linking data described below (Extended Data Fig. 5). Consistent with the crystal structure, these arrestin elements have been implicated in various aspects of arrestin activation and receptor binding[43,44].

To further characterize the rhodopsin–arrestin interfaces, we performed three additional sets of validation experiments. The first was HDX, which probes the dynamics and stability of protein complexes[45]. Compared with free arrestin, the rhodopsin-bound arrestin has several regions that are protected from exchange, including the finger loop and the N-terminal β-sheets, consistent with their location in the rhodopsin-binding interface (Fig. 4c, d and Extended Data Fig. 6a). The hydrogen to deuterium exchange rate of arrestin in the complex is lower than that for free arrestin across the whole protein, indicating that arrestin is stabilized by complex formation, consistent with the results of previous HDX experiments[46] and thermal stability assays, which revealed that the melting temperature of free arrestin is six degrees lower than that of the complex (53 °C versus 59 °C, Extended Data Fig. 6b).

The second set consisted of Tango assays[47], which have been used for probing GPCR–arrestin interactions (Extended Data Fig. 7a). Wild-type rhodopsin and wild-type arrestin had a very low basal interaction and all-*trans*-retinal increased the binding by approximately threefold. In contrast, E113$^{3.28}$Q/M257$^{6.40}$Y rhodopsin showed a high level of interaction with the pre-activated 3A arrestin, and addition of all-*trans*-retinal further increased the binding signal by approximately fivefold. Mutations in finger loop (D74, M76, G77, and L78), middle loop (Q134 and D139), and C-loop (L250 and Y251) decreased rhodopsin–arrestin binding (Extended Data Fig. 7b). Correspondingly, mutations in rhodopsin residues involved in arrestin binding also weakened the interaction (Extended Data Fig. 7c), consistent with the complex crystal structure.

The third set consisted of site-specific disulfide cross-linking experiments, which have been used to validate structures based on the geometry requirements for disulfide bond formation (Cα–Cα

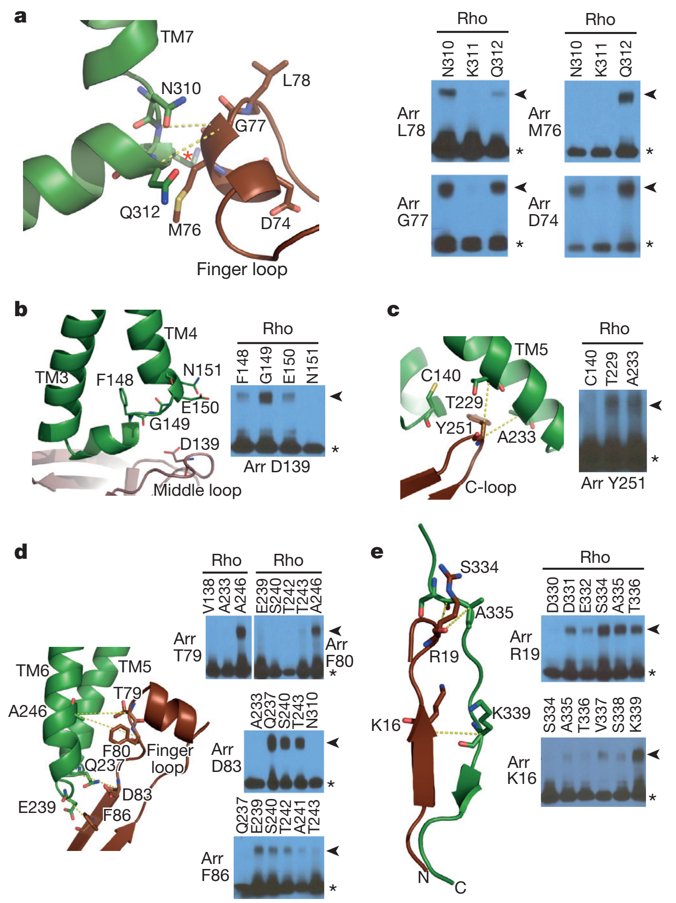


Figure 5 | Validation of the rhodopsin–arrestin interface by disulfide bond cross-linking. a–e, Structure and cross-linking of arrestin with rhodopsin. Panels are arrestin finger loop with rhodopsin TM7 and helix 8 (**a**); arrestin middle loop with rhodopsin ICL2 (**b**); arrestin C-loop residue Y251 with rhodopsin TM5 (**c**); arrestin β-strand interface residues with residues of rhodopsin TM5, ICL3, and TM6 (**d**); and arrestin's N terminus with rhodopsin's C-tail (**e**). Rhodopsin K311 is marked with a red asterisk and the side chain of arrestin M76 is shown in full from computation modelling of the full rhodopsin–arrestin complex, which was also used in panel **e**. Black asterisks, arrestin; arrowheads, rhodopsin/arrestin crosslinking adduct.

distances of 5–9Å and appropriate side-chain orientations). We engineered cysteine pairs at the binding interface of arrestin and rhodopsin, which were tagged with Flag and HA, respectively. Over 314 co-expression combinations were tested and monitored by SDS–PAGE followed by western blotting (Extended Data Fig. 8). The results are summarized in Supplementary Table 5. Every interface residue in arrestin was included in the study and the results closely agree with the crystal structure. For example, the distances from the Cα atom of the finger loop residue G77 of arrestin to the Cα atoms of N310$^{7.57}$, K311$^{8.48}$ and Q312$^{8.49}$ in rhodopsin fit the requirement for disulfide bond formation (Fig. 5a). G77C cross-linked efficiently with N310$^{7.57}$C and Q312$^{8.49}$C, but not with K311$^{8.48}$ C because the Cβ of K311$^{8.48}$ points away from G77 (Fig. 5a). Neither did G77C show cross-linking with a large set of other rhodopsin residues, indicating the high specificity of the cross-linking experiments (Extended Data Fig. 8c and Supplementary Table 5). In contrast, several other mutants in the finger loop region (D74C, M76C, and L78C) readily cross-linked with Q312$^{8.49}$C from helix 8 (Fig. 5a). The cross-linking results of these four finger loop residues not only matched the crystal structure, but also agreed well with the results from the Tango assays (Extended Data Fig. 7b). In addition, mutants of three N-terminal finger loop residues (Q70C, E71C, and D72C) were cross-linked to mutants in rhodopsin ICL1 T70$^{\mathrm{ICL1}}$C and K67$^{\mathrm{ICL1}}$C, respectively (Extended Data Fig. 7d).

We also observed cross-linking of the arrestin middle loop (D139) with rhodopsin ICL2 (G149$^{\mathrm{ICL2}}$) (Fig. 5b), of the arrestin C-loop (Y251) with rhodopsin TM5 (T229$^{5.64}$ and A233$^{5.68}$) (Fig. 5c), and of the arrestin β-strand (residues 79–86) that follows the finger loop with rhodopsin TM5, TM6, and ICL3 (Fig. 5d). Additional cross-linking was observed in two back-loop residues R319C and T320C of arrestin with Q237$^{\mathrm{ICL3}}$C of TM5 in rhodopsin (Extended Data Fig. 8e). Furthermore, extensive cross-linking of the arrestin N terminus with the C-terminal tail of rhodopsin was detected, including R19 of arrestin with S334$^{\mathrm{Cterm}}$ of rhodopsin (Fig. 5e), K16 of arrestin with S338$^{\mathrm{Cterm}}$ and K339$^{\mathrm{Cterm}}$ of rhodopsin, and V11 and S10 of arrestin with the final eight residues of rhodopsin (Supplementary Table 5). Together, these cross-linking experiments further validated the interface assembly of the rhodopsin–arrestin complex.

Possible structural mechanisms for biased signalling

The rhodopsin–arrestin complex represents the first crystal structure of a GPCR bound to arrestin and provides an opportunity to examine the mechanism of arrestin-biased signalling. Although a crystal structure of G protein-bound rhodopsin is not available, several structures of rhodopsin bound to GαCT and analogue peptides have been determined[11,22,23,39] and reveal that the arrangement of TM helices in light-activated rhodopsin is similar to that in the G-protein-bound β_2AR complex, with the exception of TM6, whose outward movement in β_2AR is much more pronounced upon binding to G protein[8]. The arrestin-bound rhodopsin has its intracellular end of TM6 moved outward by approximately 10 Å relative to its inactive structure (Fig. 6a, b and Extended Data Fig. 9). This is in contrast to the 14 Å outward movement of TM6 reported in the G-protein-bound β_2AR complex[8]. Compared to the active conformation of rhodopsin bound to GαCT peptides[11,22,23,39], arrestin-bound rhodopsin has additional conformational differences in TM1, TM4, TM5, and TM7 (Fig. 6c, d and Extended Data Fig. 9), and these unique structural features may constitute essential elements for arrestin-biased signalling.

The molecular assembly observed in the rhodopsin–arrestin complex also provides a general model for arrestin recruitment by phosphorylated rhodopsin-like class A GPCRs. In the computational model of the full complex, the highly cationic N-terminal domain of arrestin is paired with the C-terminal tail of rhodopsin (Extended Data Fig. 10). Based on the extensive disulfide crosslinking data and computation modelling, phosphorylated S334, S338 and S343 can form tight ionic interactions with three positively charged pockets at the N terminus of arrestin (Extended Data Fig. 11a–d). These results support a model of arrestin activation by phosphorylated rhodopsin through the C-tail exchange mechanism (Fig. 6e)[2]. The displacement of the arrestin C terminus by the phosphorylated rhodopsin C-tail destabilizes the polar core of arrestin[48], thus allowing for a 20° rotation of the N- and C- domains of arrestin that opens a cleft between the middle and C-loops into which the ICL2 helix of rhodopsin can fit. The ionic interaction between rhodopsin and arrestin is consistent with the fact that it is highly salt sensitive in our AlphaScreen assay (Extended Data Fig. 11e), in agreement with the salt-sensitive binding of phosphorylated rhodopsin to arrestin[43,48]. Importantly, the cytoplasmic face of the rhodopsin TM bundle is highly positively charged, whereas the finger loop (residues 70–78) contains three conserved negatively charged residues (E71, D72, and D74) (Extended Data Fig. 10). Thus, the interaction of arrestin with the rhodopsin TM bundle is mediated not only by shape but also by charge complementarity. Arrestins are highly conserved with only four subtypes in vertebrates. In contrast, there are hundreds of GPCRs, with cytoplasmic interfaces that are mainly non-conserved. However, the positive charge property is a common feature on the cytoplasmic side of a number of GPCR structures (Extended Data Fig. 12). Electrostatic interactions between arrestins and GPCRs

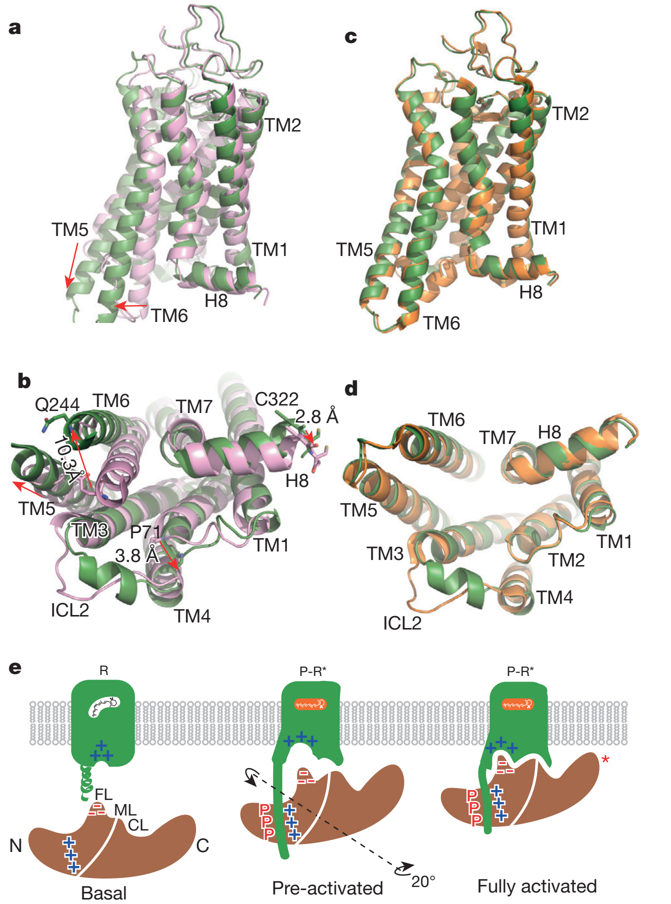


Figure 6 | Structural basis of arrestin-biased signalling and arrestin recruitment. a, **b**, Two views of structural overlays of arrestin-bound rhodopsin (green) with inactive rhodopsin (pink). **c**, **d**, Two views of structural overlays of arrestin-bound rhodopsin (green) with GαCT peptide-bound rhodopsin (orange). **e**, A cartoon model of arrestin recruitment by a phosphorylated and active rhodopsin. In the dark state, the receptor is inactive (R-state) and arrestin is in the closed state (basal state). Receptor activation and phosphorylation (P-R* state) allow the phosphorylated C-terminal tail of rhodopsin to bind to the N-domain of arrestin (pre-activated state), thus displacing the arrestin C-terminal tail. This displacement destabilizes the polar core of arrestin, which allows a 20° rotation between the arrestin N- and C-domains, leading to the opening of the middle loop (ML) and C-loop (CL) to accommodate the ICL2 helix of rhodopsin (fully activated state). The activated receptor also opens the cytoplasmic side of the TM bundle to adopt the finger loop (FL) of arrestin. In this model, the tip of arrestin's C-domain contacts the membrane (red asterisk).

may represent an adaptive mechanism for arrestins to pair promiscuously with the large number of GPCRs.

The asymmetric orientation of the bound arrestin with regard to the relative positions of its N–C domains in respect to the membrane has important implication in its binding to rhodopsin (Fig. 2a, b). Such asymmetric assembly brings the arrestin C-domain towards the membrane, with the C-edge either being touched or embedded in the membrane layer (Extended Data Fig. 13). The C-edge is comprised of conserved hydrophobic residues (F197, F198, M199, F339, and L343). It has been puzzling why single alanine mutations at these residues would affect arrestin binding to rhodopsin given how far away they are from the receptor[43]. The close proximity of these hydrophobic residues to the membrane surface may provide an explanation for the effects of these mutations on rhodopsin binding. GPCR signalling regulator proteins are normally membrane-associated through lipid modifications, as is the case for GPCR kinase 1 (GRK1) and the G-protein subunits Gα and Gβγ. Yet, there is no known lipid modification for any of the arrestins. We speculate that the conserved hydrophobic patch at the C-tip of arrestin may function as a lipid-interacting module that helps to stabilize its interaction with the receptor. Furthermore, one primary function of arrestin is to mediate endocytosis of ligand-activated GPCRs and the highly asymmetric nature of the rhodopsin–arrestin assembly may facilitate the membrane curvature for subsequent endocytotic processes. Alternatively, the remote C-tip could serve as the binding site of a second rhodopsin, which has been proposed to form dimers in the rod outer-segment disc membrane[49].

LCP-SFX is a new technology that has been used to determine several crystal structures[35,36,50]. Rhodopsin–arrestin is a challenging membrane protein complex and obtaining a structure of this complex at a sufficiently high resolution was an intractable task using existing methods that include synchrotron-based crystallography and cryo-electron microscopy[26]. The rhodopsin–arrestin complex structure reported here demonstrates the utility of X-ray lasers when combined with SFX and an LCP crystal delivery system[35]. The SFX method is relatively new and under continuous development. Given its success in solving the rhodopsin–arrestin structure, we expect that X-ray lasers, with further method development, will continue to provide breakthrough insights into biology and chemistry.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 December 2013; accepted 12 June 2015.
Published online 22 July 2015.

1. Kim, Y. J. *et al.* Crystal structure of pre-activated arrestin p44. *Nature* **497,** 142–146 (2013).
2. Shukla, A. K. *et al.* Structure of active β-arrestin-1 bound to a G-protein-coupled receptor phosphopeptide. *Nature* **497,** 137–141 (2013).
3. Pitcher, J. A., Freedman, N. J. & Lefkowitz, R. J. G protein-coupled receptor kinases. *Annu. Rev. Biochem.* **67,** 653–692 (1998).
4. Wilden, U., Hall, S. W. & Kuhn, H. Phosphodiesterase activation by photoexcited rhodopsin is quenched when rhodopsin is phosphorylated and binds the intrinsic 48-kDa protein of rod outer segments. *Proc. Natl Acad. Sci. USA* **83,** 1174–1178 (1986).
5. Reiter, E., Ahn, S., Shukla, A. K. & Lefkowitz, R. J. Molecular mechanism of beta-arrestin-biased agonism at seven-transmembrane receptors. *Annu. Rev. Pharmacol. Toxicol.* **52,** 179–197 (2012).
6. Kenakin, T. P. Biased signalling and allosteric machines: new vistas and challenges for drug discovery. *Br. J. Pharmacol.* **165,** 1659–1669 (2012).
7. Palczewski, K. *et al.* Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* **289,** 739–745 (2000).
8. Rasmussen, S. G. *et al.* Crystal structure of the β_2 adrenergic receptor-Gs protein complex. *Nature* **477,** 549–555 (2011).
9. Cherezov, V. *et al.* High-resolution crystal structure of an engineered human β_2-adrenergic G protein-coupled receptor. *Science* **318,** 1258–1265 (2007).
10. Katritch, V., Cherezov, V. & Stevens, R. C. Structure-function of the G protein-coupled receptor superfamily. *Annu. Rev. Pharmacol. Toxicol.* **53,** 531–556 (2013).
11. Standfuss, J. *et al.* The structural basis of agonist-induced activation in constitutively active rhodopsin. *Nature* **471,** 656–660 (2011).
12. Xu, F. *et al.* Structure of an agonist-bound human A2A adenosine receptor. *Science* **332,** 322–327 (2011).
13. Wang, C. *et al.* Structural basis for molecular recognition at serotonin receptors. *Science* **340,** 610–614 (2013).
14. Wacker, D. *et al.* Structural features for functional selectivity at serotonin receptors. *Science* **340,** 615–619 (2013).
15. Liu, J. J., Horst, R., Katritch, V., Stevens, R. C. & Wuthrich, K. Biased signaling pathways in β_2-adrenergic receptor characterized by ^{19}F-NMR. *Science* **335,** 1106–1110 (2012).
16. Zhou, X. E., Melcher, K. & Xu, H. E. Structure and activation of rhodopsin. *Acta Pharmacol. Sin.* **33,** 291–299 (2012).
17. Gurevich, V. V., Hanson, S. M., Song, X. F., Vishnivetskiy, S. A. & Gurevich, E. V. The functional cycle of visual arrestins in photoreceptor cells. *Prog. Retin. Eye Res.* **30,** 405–430 (2011).
18. Smith, S. O. Insights into the activation mechanism of the visual receptor rhodopsin. *Biochem. Soc. Trans.* **40,** 389–393 (2012).
19. Han, M., Smith, S. O. & Sakmar, T. P. Constitutive activation of opsin by mutation of methionine 257 on transmembrane helix 6. *Biochemistry* **37,** 8253–8261 (1998).
20. Ballesteros, J. A. & Weinstein, H. Integrated methods for the construction of three dimensional models and computational probing of structure-function relations in G-protein coupled receptors. *Methods in Neurosciences* **25,** 366–428 (1995).
21. Park, J. H., Scheerer, P., Hofmann, K. P., Choe, H. W. & Ernst, O. P. Crystal structure of the ligand-free G-protein-coupled receptor opsin. *Nature* **454,** 183–187 (2008).

22. Scheerer, P. *et al.* Crystal structure of opsin in its G-protein-interacting conformation. *Nature* **455,** 497–502 (2008).
23. Choe, H. W. *et al.* Crystal structure of metarhodopsin II. *Nature* **471,** 651–655 (2011).
24. Hirsch, J. A., Schubert, C., Gurevich, V. V. & Sigler, P. B. The 2.8 Å crystal structure of visual arrestin: a model for arrestin's regulation. *Cell* **97,** 257–269 (1999).
25. Granzin, J. *et al.* X-ray crystal structure of arrestin from bovine rod outer segments. *Nature* **391,** 918–921 (1998).
26. Shukla, A. K. *et al.* Visualization of arrestin recruitment by a G-protein-coupled receptor. *Nature* **512,** 218–222 (2014).
27. Standfuss, J., Zaitseva, E., Mahalingam, M. & Vogel, R. Structural impact of the E113Q counterion mutation on the activation and deactivation pathways of the G protein-coupled receptor rhodopsin. *J. Mol. Biol.* **380,** 145–157 (2008).
28. Xie, G., Gross, A. K. & Oprian, D. D. An opsin mutant with increased thermal stability. *Biochemistry* **42,** 1995–2001 (2003).
29. Standfuss, J. *et al.* Crystal structure of a thermally stable rhodopsin mutant. *J. Mol. Biol.* **372,** 1179–1188 (2007).
30. Martin, E. L., Rens-Domiano, S., Schatz, P. J. & Hamm, H. E. Potent peptide analogues of a G protein receptor-binding region obtained with a combinatorial library. *J. Biol. Chem.* **271,** 361–366 (1996).
31. Zhuang, T. *et al.* Involvement of distinct arrestin-1 elements in binding to different functional forms of rhodopsin. *Proc. Natl Acad. Sci. USA* **110,** 942–947 (2013).
32. Bayburt, T. H. *et al.* Monomeric rhodopsin is sufficient for normal rhodopsin kinase (GRK1) phosphorylation and arrestin-1 binding. *J. Biol. Chem.* **286,** 1420–1428 (2011).
33. Hanson, S. M. *et al.* Each rhodopsin molecule binds its own arrestin. *Proc. Natl Acad. Sci. USA* **104,** 3125–3128 (2007).
34. Boutet, S. *et al.* High-resolution protein structure determination by serial femtosecond crystallography. *Science* **337,** 362–364 (2012).
35. Weierstall, U. *et al.* Lipidic cubic phase injector facilitates membrane protein serial femtosecond crystallography. *Nature Commun.* **5,** 3309 (2014).
36. Liu, W. *et al.* Serial femtosecond crystallography of G protein-coupled receptors. *Science* **342,** 1521–1524 (2013).
37. Barty, A. *et al.* software for high-throughput reduction and analysis of serial femtosecond X-ray diffraction data. *J. Appl. Crystallogr.* **47,** 1118–1131 (2014).
38. White, T. A. *et al.* CrystFEL: a software suite for snapshot serial crystallography. *J. Appl. Crystallogr.* **45,** 335–341 (2012).
39. Deupi, X. *et al.* Stabilized G protein binding site in the structure of constitutively active metarhodopsin-II. *Proc. Natl Acad. Sci. USA* **109,** 119–124 (2012).
40. Altenbach, C., Kusnetzow, A. K., Ernst, O. P., Hofmann, K. P. & Hubbell, W. L. High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation. *Proc. Natl Acad. Sci. USA* **105,** 7439–7444 (2008).
41. Kim, M. *et al.* Conformation of receptor-bound visual arrestin. *Proc. Natl Acad. Sci. USA* **109,** 18407–18412 (2012).
42. Kirchberg, K. *et al.* Conformational dynamics of helix 8 in the GPCR rhodopsin controls arrestin activation in the desensitization process. *Proc. Natl Acad. Sci. USA* **108,** 18690–18695 (2011).
43. Ostermaier, M. K., Peterhans, C., Jaussi, R., Deupi, X. & Standfuss, J. Functional map of arrestin-1 at single amino acid resolution. *Proc. Natl Acad. Sci. USA* **111,** 1825–1830 (2014).
44. Sommer, M. E., Hofmann, K. P. & Heck, M. Distinct loops in arrestin differentially regulate ligand binding within the GPCR opsin. *Nature Commun.* **3,** 995 (2012).
45. West, G. M. *et al.* Protein conformation ensembles monitored by HDX reveal a structural rationale for abscisic acid signaling protein affinities and activities. *Structure* **21,** 229–235 (2013).
46. Ohguro, H., Palczewski, K., Walsh, K. A. & Johnson, R. S. Topographic study of arrestin using differential chemical modifications and hydrogen-deuterium exchange. *Protein Sci.* **3,** 2428–2434 (1994).
47. Barnea, G. *et al.* The genetic design of signaling cascades to record receptor activation. *Proc. Natl Acad. Sci. USA* **105,** 64–69 (2008).
48. Gurevich, V. V. & Benovic, J. L. Visual arrestin interaction with rhodopsin. Sequential multisite binding ensures strict selectivity toward light-activated phosphorylated rhodopsin. *J. Biol. Chem.* **268,** 11628–11638 (1993).
49. Fotiadis, D. *et al.* Atomic-force microscopy: rhodopsin dimers in native disc membranes. *Nature* **421,** 127–128 (2003).
50. Zhang, H. *et al.* Structure of the angiotensin receptor revealed by serial femtosecond crystallography. *Cell* **161,** 833–844 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements Portions of this research were carried out at the Linac Coherent Light Source (LCLS) at the SLAC National Accelerator Laboratory. Use of the LCLS at the SLAC National Accelerator Laboratory is supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences under contract no. DE-AC02-76SF00515. Parts of the sample injector used at LCLS for this research was funded by the National Institutes of Health, P41GM103393, formerly P41RR001209. We thank staff members of the Life Science Collaborative Access Team (ID-21) of the Advanced Photon Source (APS) for assistance in data collection at the beam lines of sector 21, which is in part funded by the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor (Grant 085P1000817), and the General Medicine Collaborative Access Team for assistance in data collection at the beam lines of sector 23 (ID-23), funded in part with Federal funds from the National Cancer Institute (ACB-12002) and the National Institute of General Medical Sciences (AGM-12006). Use of APS was supported by the Office of Science of the US Department of Energy, under contract no. DE-AC02-06CH11357. This work was supported in part by the Jay and Betty Van Andel Foundation, Ministry of Science and Technology (China) grants 2012ZX09301001 and 2012CB910403, 2013CB910600, XDB08020303, 2013ZX09507001, Amway (China), National Institute of Health grants, DK071662 (H.E.X.); GM073197 and GM103310 (C.S.P. and B.C.); GM102545 and GM104212 (K.M.); EY011500 and GM077561 (V.V.G.), EY005216 and P30 EY000331 (W.L.H.), the National Institutes of Health Common Fund in Structural Biology grants P50 GM073197 (V.C. and R.C.S.), P50 GM073210 (M.C.), and GM095583 (P.F.); National Institute of General Medical Sciences PSI: Biology grants U54 GM094618 (V.C., V.K., and R.C.S.), GM108635 (V.C.), U54 GM094599 (P.F.), GM097463 (J.S.), and U54 GM094586 (JCSG); NSF Science and Technology Center award 1231306 (J.C.H.S., P.F. and U.W.); Swiss National Science Foundation grant 31003A_141235 (J.S.); the Canada Excellence Research Chair program and the Anne & Max Tanenbaum Chair in Neuroscience at the University of Toronto (O.P.E.); and Science Foundation Ireland, grant 12/IA/1255 (M.C.). Parts of this work were also supported by the Helmholtz Gemeinschaft, the DFG Cluster of Excellence Center for Ultrafast Imaging, and the BMBF project FKZ 05K12CH1 (H.N.C., A.B., C.G., O.Y., T.W.); the Irene and Eric Simon Brain Research Foundation (R.L.). We thank A. Brunger and O. Zeldin for analysing the XFEL data and for advising on refinement; B. Weis for advice on twin refinement and structure validation; J. Rini for advice on the piggyBac expression system; A. Lebedev for his advice regarding the Zanuda program and the choice of the space group; and A. Walker for final editing of the manuscript. C.G. kindly thanks the PIER Helmholtz-Graduate School and the Helmholtz Association for financial support. We also thank the TianHe research and development team of National University of Defense Technology (NUDT) for computational resources.

Author Contributions Y.K. initiated the project, developed the expression and purification methods for rhodopsin–arrestin complex, and bulk-purified expression constructs and proteins used in LCP crystallization for the SFX method; X.E.Z. collected the synchrotron data, helped with the SFX data collection, processed the data, and solved the structures; X. Gao expressed and purified rhodopsin–arrestin complexes, characterized their binding and thermal stability, discovered the initial crystallization conditions with 9.7 MAG (1-(9Z-hexadecenoyl)-*rac*-glycerol), prepared most crystals for synchrotron data collection, prepared all crystals for the final data collection by SFX, helped with SFX data collection, and established the initial cross-linking method for the rhodopsin–arrestin complex; Y.H. designed and performed Tango assays and disulfide bond cross-linking experiments; C.Z. developed the mammalian expression methods; P.W.d.W. helped with XFEL data processing and performed computational experiments; J.K., M.H.E.T., K.M.S.-P., K.P., J.M., Y.J., X.Z., and X. Gu performed cell culture, mutagenesis, protein purification, rhodopsin–arrestin binding experiments; W.L. and A.I. grew crystals and collected synchrotron data at APS and SFX data at LCLS, G.W.H. and Q.X. determined and validated the structure. Z.Z. and V.K. constructed the full model, the phosphorylated rhodopsin–arrestin model, and helped writing the paper; D.W., S.L., D.J., C.K., Sh.B., and N.A.Z. helped with XFEL data collection and initial data analysis; Sé.B., M.M., and G.J.W. set up the XFEL experiment, performed the data collection, and commented on the paper. A.B., T.A.W., C.G., O.Y., and H.N.C. helped with XFEL data collection and data analysis, processed the data and helped with structure validation. G.M. W., B.D.P., and P.R.G. performed HDX experiments and helped with manuscript writing. J.L. helped initiate this collaborative project and with writing the paper. M.W. collected the 7.7 Å dataset at the Swiss Light Source. A.M., C.S.P., and B.C. were responsible for electron microscopy images of rhodopsin–arrestin complexes. M.T. and Y.Z. performed mass spectrometry experiments to validate the protein contents in the crystals; D.L., N. H., and M.C. provided the 9.7 MAG phase diagram and helped with SFX data collection and with writing the paper. J.S. provided a computational model of the rhodopsin–arrestin complex and helped with discussion and writing; K.D., H.L., and Y.D. helped with data analysis and twinning problems; R.J.L. constructed single-Cys arrestin-1 mutants for DEER and tested their binding to rhodopsin; S.A.V. expressed these mutants in *Escherichia coli* and purified them; V.V.G. provided arrestin genes, designed single-Cys arrestin-1 mutants for DEER, and helped analysing the data and writing the paper. H.Y. and H.J. performed computational modelling, figure preparation, and helped with writing the paper; J.C.H.S. and U.W. designed the LCP injector and helped with data collection. Sh.B., S.R.-C., C.E.C., J.C., C.K., I.G., P.F., and R.F. helped with data collection, on-site crystal characterization as well as data analysis, and validation of the structure. L.N.C. and O.P.E. generated the Y74C/C140S/C316S stable cell line, characterized and provided the rhodopsin mutant sample for DEER measurements. N.V.E. and W.L.H. incorporated rhodopsin into nanodiscs, spin-labelled rhodopsin and arrestin, performed DEER experiments and helped with manuscript writing. R.C.S. supervised crystal growth, data collection, structure solution and validation, and helped with manuscript writing. V.C. was the Principal Investigator of the LCLS data collection, supervised crystal growth, data collection at APS and LCLS, structure solution and validation, and helped with manuscript writing; K.M. supervised research, analysed data, and helped with writing the paper. H.E.X. conceived the project, designed the research, performed synchrotron and LCLS data collection and structure solution, and wrote the paper with contributions from all authors.

Author Information The coordinates of the rhodopsin–arrestin complex and diffraction data have been deposited in the Protein Data Bank under accession number 4ZWJ. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.E.X. (Eric.Xu@vai.org).

METHODS

No statistical methods were used to predetermine sample size.

Protein preparation. We used human rhodopsin and mouse visual arrestin-1 in this study. The T4L–rhodopsin–arrestin fusion protein was expressed using a tetracycline-inducible expression cassette encoding a fusion protein with His_8–MBP–MBP followed by a 3C protease cleavage site at the N terminus of the T4L–rhodopsin–arrestin. In this engineered construct, we have fused a cysteine-free T4L (residues 2–161 with C54T and C97A) to the N terminus of a rhodopsin that contains four mutations: $N2^{Nterm}C$ and $N282^{ECL3}C$ to form a disulfide bond, and $E113^{3.28}Q$ and $M257^{6.40}Y$ for constitutive receptor activity. The C terminus of rhodopsin was fused to 3A arrestin (L374A, V375A, F376A, residues 10–392) with a 15 amino acid linker (AAAGSAGSAGSAGSA).

The fusion protein constructs were expressed in HEK293S cells (Invitrogen) transiently transfected using Lipofectamine 2000 (Invitrogen). Cells were transfected at a density of 2×10^6 cells per ml at a 100 ml scale in SFM4TransFx-293 (HyClone). Six hours post-transfection, cells were diluted tenfold with CDM4HEK293 medium (HyClone). When the cell density reached approximately 4×10^6 cells per ml, protein expression was induced by the addition of doxycycline to a final concentration of $1\ \mu g\ ml^{-1}$. After 24 h induction, cells were harvested, resuspended in hypotonic buffer (20 mM HEPES-Na, pH 7.5, 10 mM NaCl, 10 mM $MgCl_2$) supplemented with EDTA-free protease inhibitor cocktail (Roche), followed by Dounce homogenization. The lysate was centrifuged at 45,000 r.p.m. at 4 °C for 1 h. The membranes were solubilized in 20 mM Tris-HCl, pH 7.4, 100 mM NaCl, 10% glycerol, 0.5% (w/v) *n*-dodecyl-β-D-maltopyranoside (DDM, Anatrace), 0.1% (w/v) cholesteryl hemisuccinate (CHS, Anatrace), and protease inhibitor cocktail for two hours at 4 °C. The supernatant was isolated by centrifugation at 45,000 r.p.m. for one hour, and incubated with amylose beads (New England Biolabs) at 4 °C overnight. Typically 10 ml of resin were used for supernatant from one litre of the original cell culture. The resin was washed with 200 ml of washing buffer (10 mM Tris-HCl, pH 7.4, 100 mM NaCl, 0.005% (w/v) lauryl maltose neopentyl glycol (MNG-3, Anatrace), 0.001% (w/v) CHS) and the protein was eluted with washing buffer containing 10 mM maltose. The eluted fusion protein was concentrated to 40–50 mg per 1 ml and the His_8–MBP–MBP tag was cleaved by overnight incubation with 3C protease at 4 °C. The cleaved His_8–MBP–MBP tag was removed by incubating with 300 µl Ni^{2+}-NTA beads (Qiagen) for three hours at 4 °C. The purified T4L–rhodopsin–arrestin was collected and concentrated to $30\ mg\ ml^{-1}$ for crystallization.

Wild-type rhodopsin and rhodopsin mutants $E113^{3.28}Q$ and $E113^{3.28}Q/M257^{6.40}Y$ used for MBP pull-down and AlphaScreen assays were expressed from the same vector with a His_8–MBP tag at the N terminus. Protein expression and purification were similar as for the rhodopsin–arrestin fusion protein used for crystallization, with the difference that proteins were not eluted, but remained bound to beads.

To generate biotinylated proteins for the AlphaScreen assays, wild-type and 3A arrestin (residues 10–392) open reading frames were cloned with N-terminal avitag–MBP tag into the first expression cassette of a modified pET-Duet expression vector (Novagen), and the biotin ligase gene *BirA* was cloned into the second cassette. The 14 amino acid avitag contains a single lysine that is efficiently biotinylated *in vivo* by the BirA protein[51]. BL21 (DE3) cells transformed with the expression plasmid were grown in LB broth at 16 °C to an OD_{600} of ~1.0 and induced with 0.1 mM IPTG in the presence of 40 µM biotin for 16 h. Cells were harvested, resuspended in 50 ml extraction buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, and 10% glycerol) per two litres of cells, and passed three times through a French Press with pressure set at 1,000 Pa. The lysate was centrifuged at 16,000 r.p.m. in a Sorvall SS-34 rotor for 30 min, and the supernatant was loaded on a 5 ml amylose HP column (GE Healthcare). The column was washed with 100 ml of wash buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, and 10% glycerol) and eluted in buffer containing 20 mM Tris-HCl, pH 8.0, 150 mM NaCl and 20 mM maltose. The eluted biotin–MBP–arrestin was concentrated and further purified by size-exclusion chromatography through a HiLoad 16/60 Superdex 200 column (GE Healthcare) in 20 mM Tris-HCl, pH 8.0, and 150 mM NaCl. Monomeric protein was collected for further assay.

Untagged 3A arrestin (residues 10–392) was expressed as a His_6–SUMO fusion protein from the expression vector pSUMO (LifeSensors). The expression and purification of untagged arrestin followed the same method as for biotin–MBP–arrestin, except that the His_6–SUMO tag was cleaved overnight with SUMO protease[52] at a protease/protein ratio of 1:1,000 in the cold room.

MBP pull-down assay. Arrestin was cloned into the pCITE-4a vector (Novagen) to allow for transcription from a T7 promoter. The TNT Quick Coupled Transcription and Translation kit was used according to the manufacturer's protocol (Promega), to express [^{35}S]methionine-labelled wild-type arrestin and 3A arrestin (L374A, V376A, and F376A). Radiolabelled wild-type arrestin and mutant arrestin proteins were incubated with His_8–MBP–rhodopsin fusion protein immobilized to 50 µl of maltose agarose bead suspension. Proteins and beads were incubated at 4 °C for one hour on a rotating platform in binding buffer containing 20 mM Tris-HCl, pH 7.4, 100 mM NaCl and 0.02% DDM/0.004% CHS. The beads were then washed three times with binding buffer and resuspended in 400 µl elution buffer (20 mM Tris-HCl, pH 7.4, 100 mM NaCl, 0.02% DDM/0.004% CHS and 100 mM maltose). The eluates were concentrated and incubated at room temperature for 15 min with 2 × loading dye. Samples were separated on 12% sodium dodecyl sulfate (SDS)-denaturing polyacrylamide gels. Gels were stained with Coomassie R-250, dried at 70 °C for 90 min, and exposed overnight to a phosphor storage screen. Results were visualized on a PhosphorImager (Fuji).

Assays for the interactions between rhodopsin and arrestin or G protein peptide. Interactions between rhodopsin and arrestin were assessed by luminescence-based AlphaScreen assay (Perkin Elmer), which our group has used extensively to determine ligand-dependent protein–protein interactions of nuclear receptors. The AlphaScreen principle is illustrated in Extended Data Fig. 1c. Briefly, biotinylated arrestin was bound to streptavidin-coated donor beads and His8-tagged rhodopsin was bound to nickel-chelated acceptor beads. The donor and acceptor beads were brought into close proximity by the interactions between rhodopsin and arrestin, which were measured in the presence or absence of all-*trans*-retinal (Sigma). When excited by a laser beam of 680 nm, the donor bead emits singlet oxygen that activates thioxene derivatives in the acceptor beads, which releases photons of 520–620 nm as the binding signal. The experiments were conducted with 100 nM of rhodopsin and arrestin proteins in the presence of $5\ \mu g\ ml^{-1}$ donor and acceptor beads in a buffer of 50 mM MOPS-Na, pH 7.4, 50 mM NaF, 50 mM CHAPS, and $0.1\ mg\ ml^{-1}$ bovine serum albumin. The results were based on an average of three experiments with standard errors typically less than 10%. GαCT-HA (TGGRVLEDLKSCGLF) and biotinylated GαCT-HA were synthesized by Peptide 2.0. For the competition assay, different amounts of untagged arrestin or GαCT-HA were added to the reaction to compete with tagged arrestin or GαCT-HA for rhodopsin binding.

Thermal stability assay. The thermal stability assay was performed with the thiol-specific fluorochrome N-[4-(7-diethylamino-4-methyl-3-coumariny)phenyl]maleimide (CPM) as described previously[53]. Briefly, 10 µg of protein was diluted with dilution buffer (20 mM Tris-HCl, pH 7.5, 200 mM NaCl, 0.005% MNG-3/CHS) to 195 µl, while CPM dye stock ($4\ mg\ ml^{-1}$ in DMSO) was freshly diluted to $0.2\ mg\ ml^{-1}$ in the dark. After 5 min of incubation at room temperature for both protein and CPM dye separately, 5 µl of diluted CPM dye was added to the protein sample and the protein/dye mix transferred immediately into a sub-micro quartz fluorometer cuvette (Starna Cells) and measured in a Cary Eclipse spectrofluorometer (Agilent). Assays were performed from 20 °C to 80 °C with a slope of 2 °C increase per minute at an excitation wavelength of 387 nm and an emission wavelength of 463 nm. All data were processed using GraphPad Prism and fit using the Boltzmann sigmoidal equation to determine the melting temperature (T_m) as inflection point of the melting curves.

Crystallization. T4L–rhodopsin–arrestin crystals were grown in lipid cubic phase (LCP)[54]. Protein solution (~$30\ mg\ ml^{-1}$) was mixed with monopalmitolein (9.7 MAG, from Nu-Chek Prep, Inc.) containing 10% cholesterol at a 1:1 ratio (w/w) using a coupled syringe mixer[55] and 50 nl boluses of protein-laden LCP were dispensed on 96-well glass sandwich plates (Molecular Dimensions or Marienfeld-Superior) and overlaid with 0.8 µl precipitant solutions using a Gryphon LCP robot (Art Robbins Instruments) or an NT8-LCP robot (Formulatrix). Multiple initial hits were identified by using screens of 30% PEG 400 in combination of 100 mM or 400 mM salts from the StockOptions Salt kit (Hampton Research)[56]. Crystals that reached full size (around 10–20 µm) within four days at 20 °C were harvested from the mesophase and were flash frozen in liquid nitrogen without additional cryoprotectant. Crystals used for synchrotron data collection were grown in 0.05 M magnesium acetate, 0.05 M sodium acetate, pH 5.0 and 28% PEG 400.

Crystals for LCP-SFX were prepared in 100 µl gas-tight Hamilton syringes as described[36,57]. About 5 µl of protein-laden LCP in the presence of fivefold molar excess of all-*trans*-retinal was slowly injected into 60 µl mother liquid buffer (0.15 M ammonium phosphate, pH 6.4 and 32% PEG 400) using a coupled syringe mixing device[55]. Crystals were grown in several syringes at 20 °C, consolidated and transferred into the LCP injector[35] for XFEL diffraction data collection. The average crystal size was 5–10 µm as determined under a polarized light microscope. The phase diagram for 9.7 MAG suggests that this MAG is a suitable host lipid for extruding LCP in vacuum at 20 °C, where evaporative cooling created problems when 9.9 MAG was used as a host LCP lipid[35].

Data collection (synchrotron). A partial 8.0 Å synchrotron data set was collected at 100 K using an X-ray beam at the wavelength of 1.0 Å at the 21 ID-D beam line of LS-CAT and at 23-ID-D of GM-CAT at the Advanced Photon Source at

Argonne National Laboratory. A full 7.7 Å data set was collected from a single crystal (~20 μm in size) using 10 μm beam size and 0.1 s exposures per 0.1° oscillation with a Pilatus 6M pixel detector at the X10SA beam line at the Swiss Light Source. The observed reflections were reduced, merged, and scaled with XDS[58] with statistics shown in Supplementary Table 1a. The L-test plot of the 7.7 Å data set is consistent with a perfectly twinned crystal.

Data collection (XFEL). LCP-SFX experiments were carried out at the Coherent X-ray Imaging (CXI) instrument[59] at the Linac Coherent Light Source (LCLS) in the SLAC National Accelerator Laboratory (Menlo Park, California, USA). X-ray pulses of 50 fs duration at a wavelength of 1.3 Å (9.5 keV) were attenuated to ~3% (3×10^{10} photons per pulse) and focused to ~1.5 μm diameter at the interaction point using Kirkpatrick–Baez mirrors[60]. Rhodopsin–arrestin complex crystals in LCP were injected across the XFEL beam using an LCP injector[35] with a 50 μm diameter nozzle at a flow rate of ~0.2 μl min^{-1}. Diffraction patterns were collected at 120 Hz using the Cornell-SLAC Pixel Array Detector (CSPAD). Over 5 million data frames were collected corresponding to ~12 h of data acquisition time. Of these frames, ~0.45% images contained potential crystal hits as identified using Cheetah[61] (more than 40 Bragg peaks of 1–20 pixels in size and a signal to noise ratio better than 6 after local background subtraction). Of the potential crystal hits, 18,874 diffraction patterns could be auto-indexed by CrystFEL[38] using a combination of MOSFLM[62], XDS[58] and DirAx[63]. An integration radius of only two pixels was used to avoid overlapping with neighbouring peaks due to the high spot density resulting from the large unit cell dimensions. Partial reflections from different crystals in random orientations were merged using a Monte Carlo integration across the crystal rocking curve of each reflection[64]. The resolution was anisotropic with ~3.3 Å resolution along the *c**-axis and ~3.8 Å resolution along the *a**/*b** axes. The data used for the structure refinement were truncated at 3.8 Å/3.8 Å/3.3 Å using the get_hkl program of CrystFEL[38] based on the criteria of data correlation coefficient (CC*), which is 0.87 at the highest resolution shell (Supplementary Table 1a). The use of CC* of 0.5 as resolution cutoff has been recently recommended for X-ray diffraction[65]. The resolution cutoff for several published XFEL structures follows this criterion, including the $5HT_{2B}$ GPCR XFEL structure[36], which has a CC* of 0.74. The statistics of the final data used in structure refinement are shown in Supplementary Table 1b.

Structure determination. The XFEL data were initially merged according to the apparent Laue group of 4/mmm, and molecular replacement searches were performed in all possible space groups of 4/m and 4/mmm. The best structure solution was found in $P4_3$. Based on analysis of the Zanuda program[66], we determined that the most likely space group was $P2_12_12_1$ and that the crystals were physically twinned. The data were reprocessed with the Laue group of mmm and molecular replacement searches were performed in $P2_12_12_1$. This space group assignment resulted in the best statistics and map quality out of several possible space groups (Supplementary Table 2).

The crystals appeared to be pseudo-merohedrally twinned based on L-test analysis[67]. Despite the challenge of twinned data, the rhodopsin–arrestin complex structure was solved by the molecular replacement method implemented in Phaser[68] using the models of constitutively active rhodopsin, pre-activated arrestin, and T4L (PDB codes: 4A4M[39], 4J2Q[1], and 3SN6[8], respectively). Four molecules of rhodopsin and four molecules of arrestin were found sequentially by molecular replacement search, resulting in four very similar rhodopsin–arrestin assemblies. Four T4Ls were also found in the aqueous layer with its C-terminal residue in a position to form a covalent bond with the first residue of rhodopsin, supporting the correct positioning of T4Ls by molecular replacement.

The structure was initially refined against the XFEL data without twin law to an R_{free}-factor of ~36% and the model maps from the data were of sufficient quality to interpret the overall structure of the rhodopsin–arrestin complex (Extended Data Fig. 3). The model then underwent iterated cycles of manual building into $2F_o - F_c$ maps with Coot[69] and refinement with REFMAC[70] and the PHENIX[71], where rigid body, individual position, group B-factor, and TLS refinements were used along with NCS restraints and twin law (k, h, -l). The arrestin residues 70–78 and 165–175, which were not included in the molecular replacement model, were manually placed into the density map. These two regions of arrestin became visible because they are either engaged in direct interaction with rhodopsin or involved in crystal packing. Regions with poor density were removed from the final model, including T4L from the B complex and the N-domain of T4L (residues 13–57) from the C complex. We did not observe clear electron density for the all-*trans*-retinal, which was thus not included in the structure. The structure has been carefully refined to the final state that has excellent geometry and refinement statistics (Supplementary Table 1b). Ramachandran plot analysis indicates that 100% of the residues are in favourable or allowed regions (no outliers). The final structure was validated with MolProbity, which revealed an all-atom-clash score of 1.47 and MolProbity score of 1.13[72].

The real-space correlation coefficients against a $2mF_o - DF_c$ map for each chain of the structure on a per residue basis using the CCP4 EDSTATS program[73] or the MolProbity program in Phenix indicated an overall good fit between the structure and the electron density map. The density fit correlation in Coot was low with the structure from Phenix twinned refinement because the map from the MTZ file with map coefficients produced by phenix.refine was not in absolute scale. This problem was overcome by using the $2F_o - F_c$ CNS format map from phenix.refine, which generated a normal correlation in Coot, similar to those from EDSTATS and MolProbity. All structural figures were prepared using PyMOL[74].

Cell-based assays for rhodopsin–arrestin interactions (Tango assays). For the cell-based Tango assay[47], we generated fusion constructs consisting of rhodopsin (1-321), a tobacco etch virus (TEV) protease cleavage site (TEV site), and the transcriptional activator tTA (Rho–TEV site–rTA) as well as of arrestin (10–392) and TEV protease (Arr–TEV protease). HTL cells were a gift from G. Barnea and R. Axel (Brown University and Columbia University). 10 ng Rho–TEV site–tTA construct were transfected together with 10 ng Arr–TEV protease plasmid and 1 ng of phRG-tk *Renilla* luciferase expression vector into HTL cells using Xtremgene (Roche). One day after transfection, cells were induced by vehicle (DMSO) or all-*trans*-retinal (10 μM) overnight. Cells were harvested and lysed in Passive Lysis Buffer (Promega). Luciferase activity was measured using the Dual Luciferase Kit (Promega) according to manufacturer's instructions.

Mutagenesis. Site-directed mutagenesis was carried out using the QuikChange method (Agilent). Mutations and all plasmid constructs were confirmed by sequencing before protein expression, MBP-pull down assay, Tango assay, and AlphaScreen assay.

Electron microscopy studies of the rhodopsin–arrestin complex. Samples were prepared as previously described[75]. Briefly, the sample was applied to a freshly glow-discharged carbon coated copper grid and allowed to adhere for 10 s before being reduced to a thin film by blotting. Immediately after blotting 3 μl of a 1% solution of uranyl formate was applied to the grid and blotted off directly. This was repeated three times. Data were acquired using a Tecnai F20 Twin transmission electron microscope operating at 200 kV, with a dose of ~40 $e^- \text{Å}^{-2}$ and nominal underfocus ranging from 2 to 3 μm. Images were automatically collected at a nominal magnification of 62,000 × and pixel size of 0.273 nm. All images were recorded with a Tietz F416 4k × 4k pixel CCD camera using Leginon data collection software[76]. Experimental data were processed by the Appion software package[77], which interfaces with the Leginon database infrastructure. ~6,000 particles were automatically extracted from 54 electron microscopy micrographs[78] and the particle stack was then aligned and sorted using the XMIPP reference-free maximum likelihood alignment[79]. Several exemplary 2D averages are shown in Fig. 1e, which were derived from class averages for the complex with or without T4L, computed from ~14,000 particles selected from 155 electron microscopy-micrographs.

In-cell disulfide bond cross-linking. The open reading frames of full-length arrestin with C-terminal Flag tag and full-length rhodopsin with C-terminal haemagglutinin (HA) tag were cloned into pcDNA6. Cysteine mutations (41 for arrestin and 51 for rhodopsin) were systematically introduced into arrestin and rhodopsin in these two DNA vectors. AD293 cells were split one day before transfection at 50,000 cells per well in a 24-well plate. Cells were grown for one day, then transfected with 100 ng rhodopsin constructs (pcDNA6-rho-3HA) plus 100 ng arrestin plasmid (pcDNA6-Arr-3Flag) by Lipofectamine 2000 (DNA/Lipofectamine 2000 ratio of 1:2) in each well. Cells were grown for 2 days after transfection, and were then treated at room temperature with H_2O_2, which was freshly diluted in the cell culture medium to a final concentration of 1 mM. After 5 min treatment with H_2O_2, the medium was aspirated and 100 μl of CelLytic M (Sigma C2978) were added to each well and the plate was shaken for 10 min at room temperature. Cell lysates were transferred to 1.5 ml tubes and spun at 16,000*g* at 4 °C for 5 min. The supernatants (10 μl) were mixed with an equal volume of 2 × SDS loading buffer (without reducing agents) for 5 min at room temperature, and loaded onto a protein gel for western blot analysis. Horseradish peroxidase-conjugated anti-Flag (Sigma M2) and anti-HA (Sigma) antibodies were used to probe for free and cross-linked arrestin and rhodopsin proteins.

Hydrogen-deuterium exchange mass spectrometry (HDX). HDX was carried out as described previously[80], with the following modifications: (1) the solution handling and mixing was performed with a LEAP Technologies Twin HTS PAL liquid handling robot housed inside a temperature-controlled cabinet held at 4 °C and (2) decyl maltose neopentyl glycol (DMNG) was used in place of DDM in the exchange buffer. Briefly, all stock solutions and dilutions were made using the 7TM HDX buffer (50 mM HEPES (pH 7.5), 150 mM NaCl, 2% (v/v) glycerol, 0.01% (m/v) CHS and 0.05% (m/v) DMNG in either H_2O or in D_2O for on-exchange). All HDX protein stock solutions were prepared at 15 μM in the 7TM

HDX H_2O buffer. On-exchange was carried out in triplicate for predetermined times (10, 30, 60, 900 and 3,600 s) at 4 °C by mixing 5 μl of stock protein solution with 20 μl of D_2O on-exchange buffer. Exchange was quenched by adding 25 μl of quench solution (100 mM NaH_2PO_4, 0.02% DMNG, and 15 mM TCEP at pH 2.4) to the reaction. Digestion was performed in line with chromatography using an in-house packed pepsin column. Peptides were captured and desalted on a C8 trap. Peptides were then separated across a 5μ 10 × 1 mm Betasil C8 column (Thermo Fisher Scientific) with a linear gradient of 12–40% acetonitrile in 0.3% formic acid over a short 5 min gradient to limit back exchange with the solvent.

Mass spectra were acquired in the range of *m/z* 300–2,000 at a resolution of 60,000 for 8 min in positive ion mode on a Q Exactive mass spectrometer (Thermo Fisher Scientific) equipped with an ESI source operated at a capillary temperature of 225 °C and spray voltage of 3.5 kV. The intensity weighted average m/z value (centroid) of each peptide's isotopic envelope was calculated with the Workbench program[81] and converted to % deuterium values. Back-exchange correction was based on an estimated 70% deuterium recovery and accounting for the known 80% deuterium content of the on-exchange buffer. The Workbench software used *P* values lower than 0.05 for two consecutive time points to determine significance. Sequence coverage experiments for these proteins were carried out in the 7TM HDX buffer and LC system described above, but with a longer 60 min gradient. 75 pmol of protein were loaded onto the column. For sequencing, tandem mass spectra were obtained using data-dependent acquisition with 30 s dynamic exclusion, where the top five most abundant ions in each scan were selected and subjected to CID fragmentation. Each scan was the average of 3 microscans under normal scan mode in both MS and MS/MS. Peptides were identified by searching spectra against an in-house database using the mascot search engine as described previously[45].

Generation of a stable cell line expressing bovine rhodopsin mutant Y74C/C140S/C316S and rhodopsin preparation. The rhodopsin mutant Y74C/C140S/C316S gene[82] was PCR-amplified from the pMT vector by using two flanking primers containing NheI and NotI restriction sites and subcloned into vector PB-T-PAF[83]. The insert in the resulting plasmid termed PB-Rho-Y74C was verified by automated DNA sequencing.

Generation of the stable cell line as well as rhodopsin production and purification was performed as described[84]. HEK293S GnTI$^-$ cells[85] were cultured in DMEM/F12 (Wisent) supplemented with 10% heat inactivated FBS (Life Technologies) and 100 U ml^{-1} penicillin/streptomycin (Life Technologies) and incubated at 37 °C, 5% CO_2. For transfection, cells were plated in 6-well plates at a density of ~600,000 cells per well. The following day, cells were co-transfected with 4 μg of PB-Rho-Y74C, 0.5 μg of PB-RB[83] and 0.5 μg of pCyL43[83,86] using JetPrime (PolyPlus, France) reagent following the manufacturer's instructions and medium was replaced by fresh medium 4 h later. (Plasmids PB-T-PAF and PB-RB were provided by James Rini, University of Toronto, Canada, and pCyL43 by the Wellcome Trust Sanger Institute, UK). Two days after transfection, cells were transferred to a 100-mm tissue culture plate. Dual drug selection was started the following day, using 10 μg ml^{-1} puromycin (Bioshop, Canada) and 5 μg ml^{-1} blasticidin (Bioshop, Canada) and lasted for 2 weeks. Cells were then transferred to larger flasks and finally to pleated roller bottles (Thermo Scientific, USA) containing 250 ml of DMEM/F12 supplemented with 10% heat-inactivated FBS and 100 U ml^{-1} of penicillin and 100 μg ml^{-1} streptomycin. Roller bottles were incubated at 37 °C, 5% CO_2 and rotated at 0.1 r.p.m. Expression was induced after 5–6 days by replacing medium with 250 ml of fresh medium containing 1 μg ml^{-1} doxycycline (Bio Basic) and 1 μg ml^{-1} aprotinin (Bioshop Canada). On day 3 after induction, cells were harvested, using PBS-EDTA with protease inhibitor tablet (Roche) to detach cells. Cell pellets were flash frozen in liquid nitrogen and stored at −80 °C until purification.

For purification, cell pellets with rhodopsin mutant Y74C/C140S/C316S were incubated with 11-*cis* retinal to reconstitute rhodopsin which was purified and spin labelled with 1-oxyl-2,2,5,5-tetramethyl-3-pyrroline-3-methyl methanethiosulfonate (MTSSL) as previously described[41].

Incorporation of spin labelled rhodopsin into nanodiscs. Nanodiscs were prepared by mixing cholate solubilized lipid (70% POPC + 30% POPS), scaffolding protein MSP1E3D1[87], and rhodopsin (in 90 mM β-OG) in a molar ratio of 140:1:0.1. The OG and cholate were removed from the mixture by dialysis against buffer D (20 mM MOPS, 150 mM NaCl, pH 6.8). The nanodiscs were further purified by passing them over a nickel NTA column to remove any receptor not incorporated into the discs. Upon elution from the nickel column with imidazole, the nanodiscs were buffer exchanged into buffer D containing 10% glycerol.

Spin labelling of arrestin mutants. C-terminally truncated mutants of arrestin, lacking endogenous cysteines, have been shown to bind to non-phosphorylated rhodopsin[88]. Single-cysteine mutants in this background were expressed in *E. coli*, as described[89], spin labelled in buffer D overnight using a tenfold molar excess of MTSSL. The base mutant also contained two alanine substitutions (F85A and F197A) which have been shown to disrupt arrestin-1 oligomerization[89]. Non-covalently bound spin label was removed from the sample by extensive washes with buffer D using Amicon 10 kDa concentrators.

DEER spectroscopy of the rhodopsin–arrestin complex. Preparations of rhodopsin (Y74$^{2.41}$C) and bovine arrestin (S60C, V139C, and L240C, which correspond to the mouse arrestin T61C, V140C, and S241C) and their spin labelling were performed as described previously[41]. For DEER measurements, the spin-labelled proteins were mixed in 1:1 ratio in the dark and loaded into quartz capillaries (1.5 mm internal diameter and 1.8 mm outer diameter). The samples were irradiated for 30 s within the capillaries using a tungsten light source with a 500 nm cutoff filter. Immediately after irradiation, the samples were flash frozen in liquid nitrogen, and loaded into an EN 5107D2 resonator for Q band DEER measurements. Measurements were performed at 80 K on a Bruker Elexsys 580 spectrometer with a Super Q-FTu Bridge. A 36-ns π-pump pulse was applied to the low field peak of the nitroxide field swept spectrum, and the observer π/2 (16 ns) and π (32 ns) pulses were positioned 50 MHz (17.8 G) upfield, which corresponds to the nitroxide centre line. Model-free distance distributions were obtained from the raw dipolar evolution data using the LabVIEW (National Instruments) program "LongDistances" that can be downloaded from http://www.biochemistry.ucla.edu/biochem/Faculty/Hubbell/.

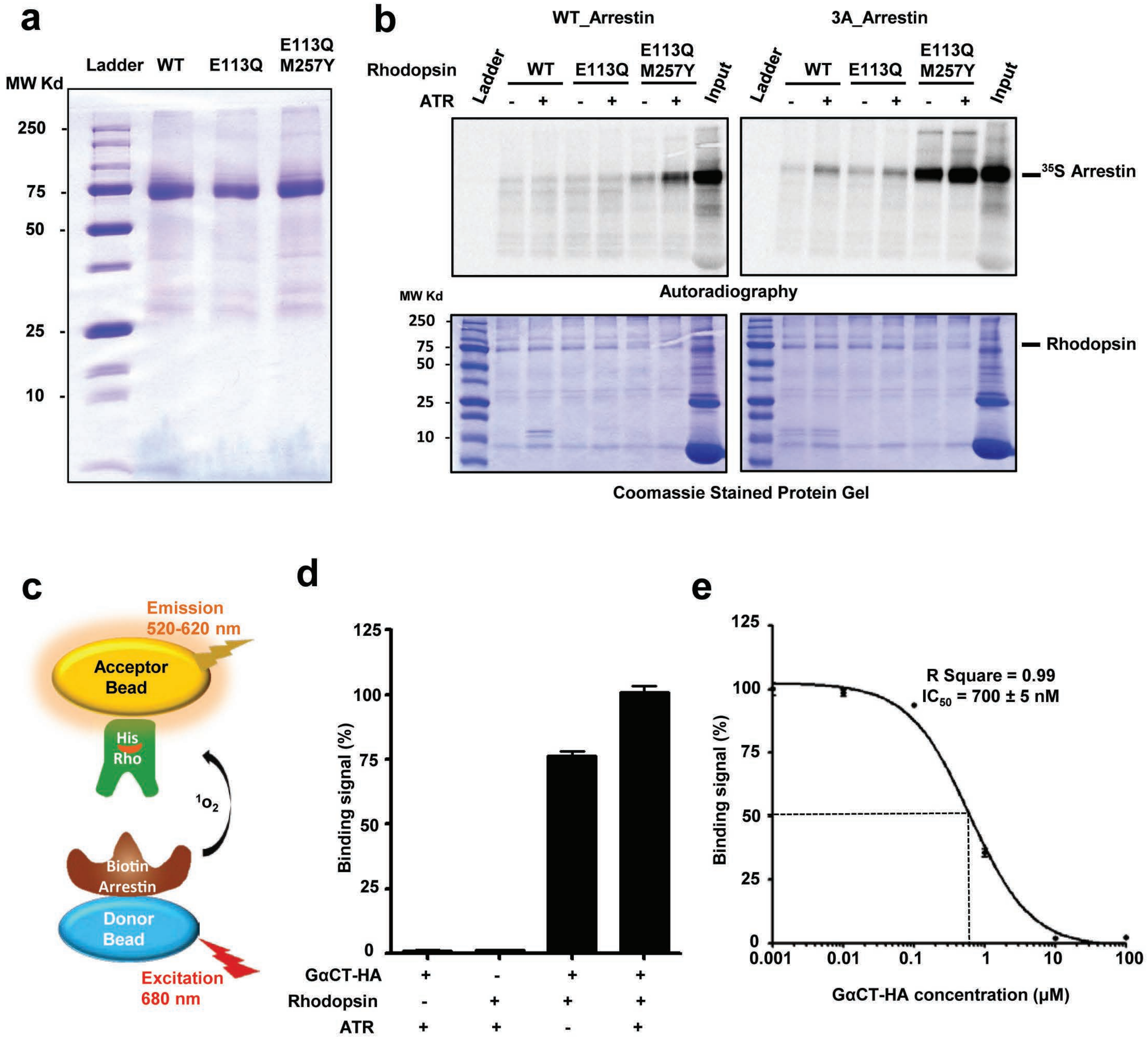
To estimate the median distances, the distance distributions were integrated and normalized to the maximum amplitude. The median distance was estimated as that corresponding to 0.5 of the integrated intensity. The modelled distances between nitroxide spin labels are based on the crystal structure of the rhodopsin–arrestin complex. R1 nitroxide side chains were modelled into the structure using common R1 rotamers[90,91].

9.7 MAG/water temperature-composition phase diagram. The phase diagram was constructed based on small- and wide-angle X-ray scattering measurements made in the heating direction. Sample preparation and X-ray scattering measurements and analysis were as previously described[55,92]. The phases identified include the lamellar crystalline (Lc) or solid phase, the fluid isotropic (FI) or liquid phase, and the following liquid crystalline phases: lamellar liquid crystal (L_α), cubic-Ia3d and cubic-Pn3m. A separate aqueous phase observed in equilibrium with the solid or liquid crystalline phases is indicated by Aq. The phase diagram shows that the solid Lc phase stabilizes under equilibrium conditions below ~8 °C. The latter is some 10 °C below that observed with 9.9 MAG (monoolein)[92] and is similar to what was found with 7.9 MAG[93]. This low solidification temperature enabled use in the current project of 9.7 MAG as the host lipid for LCP-SFX data collection in an evacuated sample chamber at 20 °C, where evaporative cooling created problems for measurements with 9.9 MAG but not with 7.9 MAG[35]. The maximum water carrying capacity of 9.7 MAG resides at ~50%(w/w) water, which is considerably greater and smaller than that of 9.9 MAG[92] and 7.7 MAG[94], respectively. These observations indicate that the cubic mesophase of 9.7 MAG has larger aqueous channels compared to 9.9 MAG that are more like those of 7.7 MAG. This is consistent with 9.7 MAG supporting the growth of rhodopsin–arrestin–T4L crystals where the complex has sizable extra-membrane bulk best accommodated in a large aqueous channel. This parallels the rational use of 7.7 MAG as a host lipid for the β_2AR–Gs complex crystallization and structure determination[8].

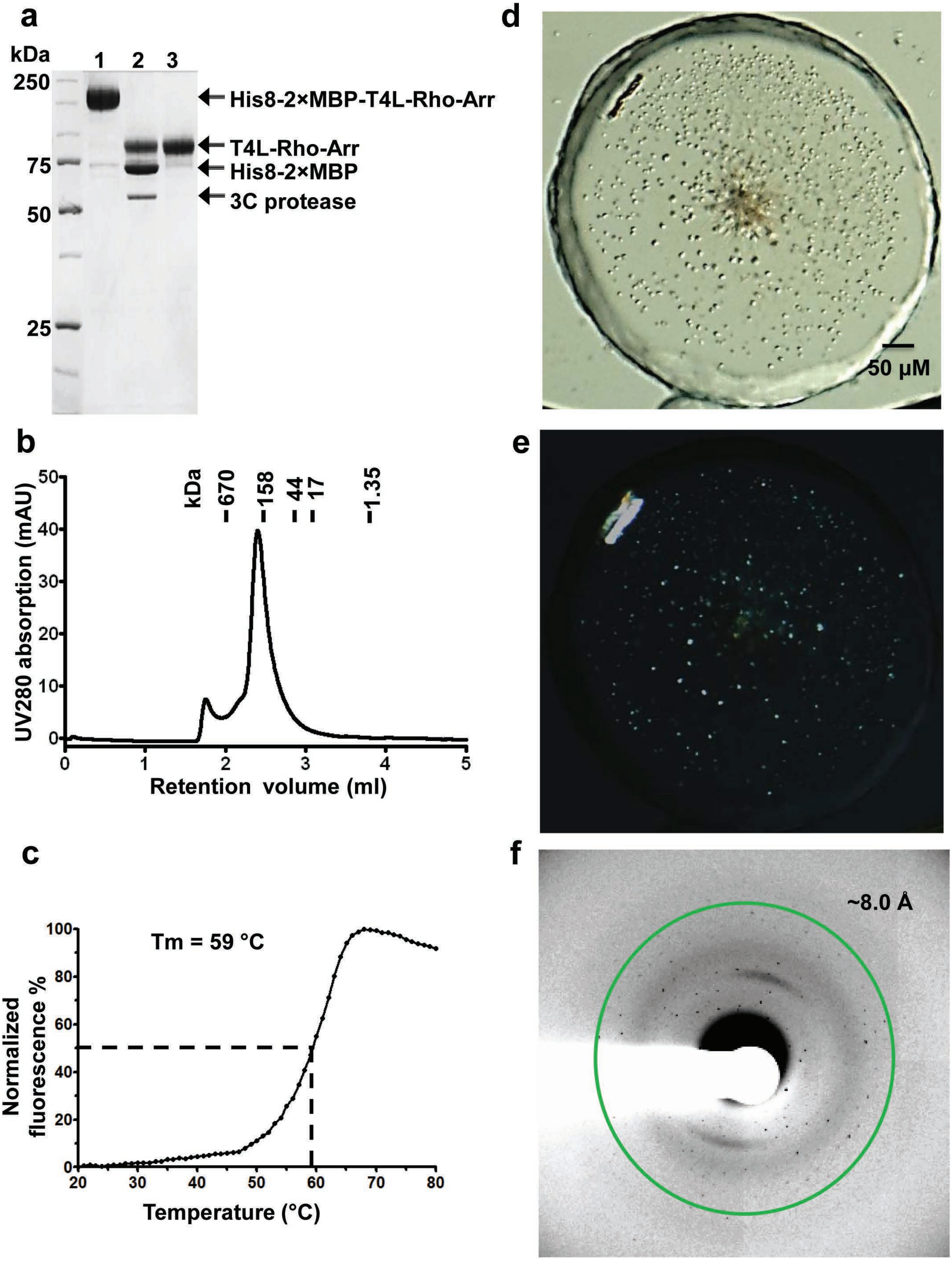
Molecular modelling of the full-length rhodopsin–arrestin complex. Energy-based conformational modelling of the rhodopsin–arrestin complex was performed with the ICM-Pro molecular suite[95], using a global energy optimization procedure similar to the one described recently for modelling of the full-length complex of CRFR1[96]. Protein sequences of human rhodopsin and mouse arrestin were obtained from the Uniprot database (http://www.uniprot.org/). Starting from the crystallographically determined structure of the complex, the modelling procedure was used to add unresolved residues of the C terminus (residues 327–345) of the human rhodopsin structure, as well as missing residues in the arrestin loop (residues 340–342) of the mouse arrestin structure. The final model did not include the last 3 residues of rhodopsin (346–348), which lack well-defined cross-linking contacts and appear flexible. Initial conformations of the short loop in arrestin were predicted with the fast "build model" ICM algorithm, followed by extensive energy optimization in internal coordinates. Conformational optimization of the rhodopsin C-terminal peptide was guided by soft pairwise harmonic distance restraints derived from disulfide crosslinking data. The restraints introduced between Cβ atoms of the crosslinked residues were graded according to the crosslinking strength listed in Supplementary Table 5, from very strong, with the penalty function starting at 5 Å distance, to medium at 7 Å distance and very weak at 12 Å. The C-terminal peptide conformation and conformation of the contact side chains of the arrestin were optimized to convergence (3 independent simulations of 10^6 steps) using global optimization procedure in internal coordinates with improved conformational energy terms for protein and peptides[97]. A special backbone closure sampling procedure was applied to the loop regions to allow efficient optimization. The global optimization runs were executed in parallel on a Linux multicore server resulting in similar best energy conformations (<3 Å r.m.s.d.; root mean squared deviation) for the C-terminal peptide residues.

The best energy-optimized conformation of this region suggests that the extended C terminus peptide runs antiparallel along the N-terminal β-strand of arrestin. This conformation of the C terminus satisfied all 17 medium to strong disulfide crosslinking restraints for this region (Supplementary Table 5), while making a number of specific polar interactions and salt bridges of D331, E332 and E341 side chains with arrestin basic residues (Extended Data Fig. 5). These modelling results suggest that although the rhodopsin C terminus is rather flexible, some low energy conformations may be preferable even in non-phosphorylated rhodopsin. Moreover, independent modelling of the complex with phosphorylated serine residues Ser334, Ser338 and Ser343 in the C terminus of rhodopsin resulted in a similar conformation of this domain. The interactions within phosphorylated complex, however, are greatly enhanced by as many as seven additional salt bridges between negatively charged phosphates and the positively charged lysine and arginine residues within the N- terminal domain of arrestin (Extended Data Fig. 11).

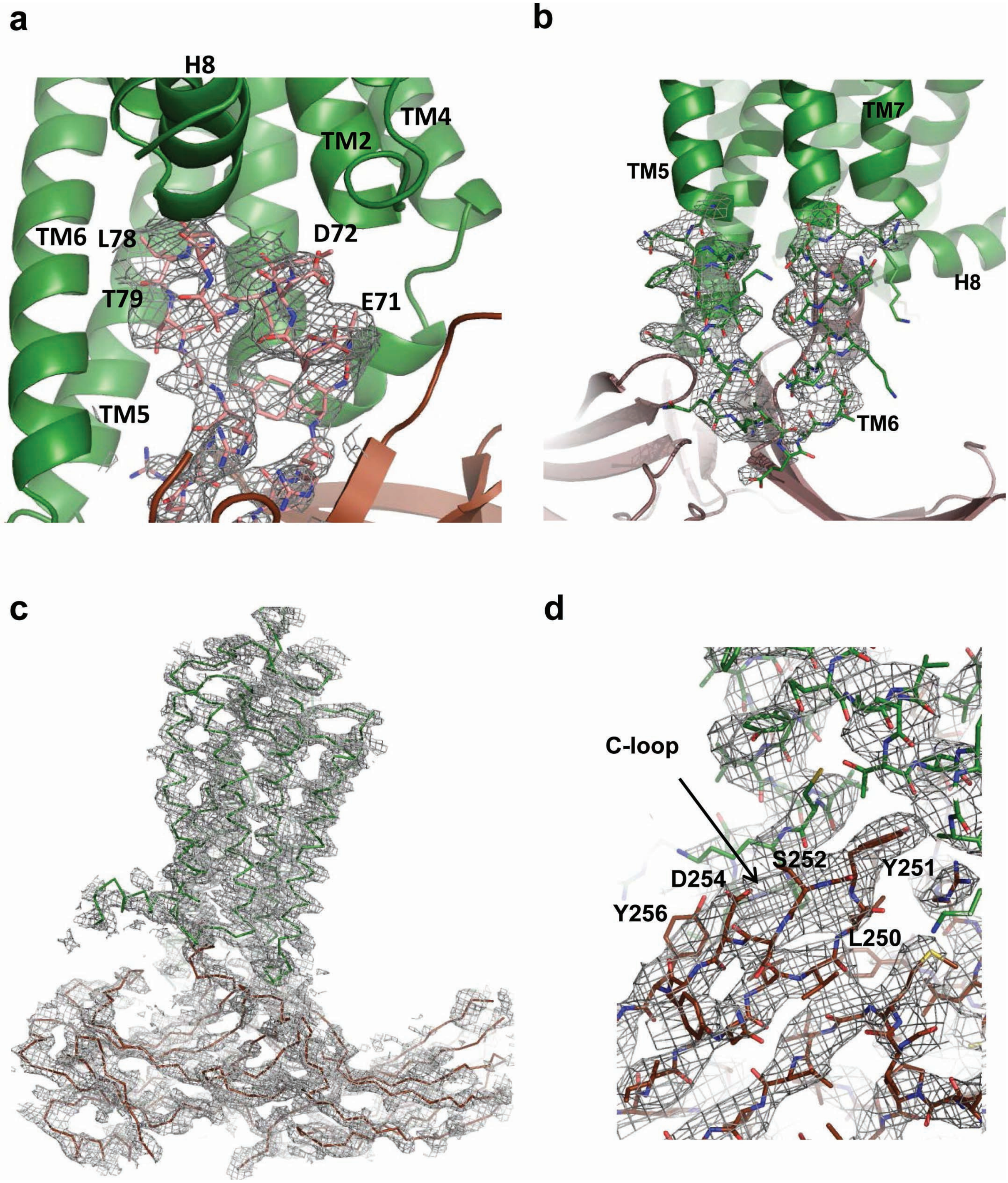
51. Beckett, D., Kovaleva, E. & Schatz, P. J. A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. *Protein Sci.* **8,** 921–929 (1999).
52. Mossessova, E. & Lima, C. D. Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast. *Mol. Cell* **5,** 865–876 (2000).
53. Alexandrov, A. I., Mileni, M., Chien, E. Y. T., Hanson, M. A. & Stevens, R. C. Microscale fluorescent thermal stability assay for membrane proteins. *Structure* **16,** 351–359 (2008).
54. Caffrey, M. & Cherezov, V. Crystallizing membrane proteins using lipidic mesophases. *Nature Protocols* **4,** 706–731 (2009).
55. Chen, A. H., Hummel, B., Qiu, H. & Caffrey, M. A simple mechanical mixer for small viscous lipid-containing samples. *Chem. Phys. Lipids* **95,** 11–21 (1998).
56. Xu, F., Liu, W., Hanson, M. A., Stevens, R. C. & Cherezov, V. Development of an automated high throughput LCP-FRAP assay to guide membrane protein crystallization in lipid mesophases. *Cryst. Growth Des.* **11,** 1193–1201 (2011).
57. Liu, W., Ishchenko, A. & Cherezov, V. Preparation of microcrystals in lipidic cubic phase for serial femtosecond crystallography. *Nature Protocols* **9,** 2123–2134 (2014).
58. Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr. D* **66,** 133–144 (2010).
59. Boutet, S. & Williams, G. J. The Coherent X-ray Imaging (CXI) instrument at the Linac Coherent Light Source (LCLS). *New J. Phys.* **12,** 035024 (2010).
60. Siewert, F. *et al.* Ultra-precise characterization of LCLS hard X-ray focusing mirrors by high resolution slope measuring deflectometry. *Opt. Express* **20,** 4525–4536 (2012).
61. White, T. A. *et al.* Crystallographic data processing for free-electron laser sources. *Acta Crystallogr. D* **69,** 1231–1240 (2013).
62. Battye, T. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr. D* **67,** 271–281 (2011).
63. Duisenberg, A. J. M. Indexing in single-crystal diffractometry with an obstinate list of reflections. *J. Appl. Crystallogr.* **25,** 92–96 (1992).
64. Kirian, R. A. *et al.* Structure-factor analysis of femtosecond microdiffraction patterns from protein nanocrystals. *Acta Crystallogr. A* **67,** 131–140 (2011).
65. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science* **336,** 1030–1033 (2012).
66. Lebedev, A. A. & Isupov, M. N. Space-group and origin ambiguity in macromolecular structures with pseudo-symmetry and its treatment with the program Zanuda. *Acta Crystallogr. D* **70,** 2430–2443 (2014).
67. Padilla, J. E. & Yeates, T. O. A statistic for local intensity differences: robustness to anisotropy and pseudo-centering and utility for detecting twinning. *Acta Crystallogr. D* **59,** 1124–1130 (2003).
68. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
69. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
70. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D* **67,** 355–367 (2011).
71. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
72. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
73. Tickle, I. J. Statistical quality indicators for electron-density maps. *Acta Crystallogr. D* **68,** 454–467 (2012).
74. DeLano, W. L. & Lam, J. W. PyMOL: a communications tool for computational models. *Abstr. Pap. Am. Chem. Soc.* **230,** U1371–U1372 (2005).
75. Moeller, A., Kirchdoerfer, R. N., Potter, C. S., Carragher, B. & Wilson, I. A. Organization of the influenza virus replication machinery. *Science* **338,** 1631–1634 (2012).
76. Suloway, C. *et al.* Automated molecular microscopy: the new Leginon system. *J. Struct. Biol.* **151,** 41–60 (2005).
77. Lander, G. C. *et al.* Appion: An integrated, database-driven pipeline to facilitate EM image processing. *J. Struct. Biol.* **166,** 95–102 (2009).
78. Voss, N. R., Yoshioka, C. K., Radermacher, M., Potter, C. S. & Carragher, B. DoG Picker and TiltPicker: Software tools to facilitate particle selection in single particle electron microscopy. *J. Struct. Biol.* **166,** 205–213 (2009).
79. Scheres, S. H. W., Nunez-Ramirez, R., Sorzano, C. O. S., Carazo, J. M. & Marabini, R. Image processing for electron microscopy single-particle analysis using XMIPP. *Nature Protocols* **3,** 977–990 (2008).
80. Goswami, D. *et al.* Time window expansion for HDX analysis of an intrinsically disordered protein. *J. Am. Soc. Mass Spectrom.* **24,** 1584–1592 (2013).
81. Pascal, B. D. *et al.* HDX Workbench: software for the analysis of H/D exchange MS data. *J. Am. Soc. Mass Spectrom.* **23,** 1512–1521 (2012).
82. Klein-Seetharaman, J. *et al.* Single-cysteine substitution mutants at amino acid positions 55-75, the sequence connecting the cytoplasmic ends of helices I and II in rhodopsin: reactivity of the sulfhydryl groups and their derivatives identifies a tertiary structure that changes upon light-activation. *Biochemistry* **38,** 7938–7944 (1999).
83. Li, Z., Michael, I. P., Zhou, D., Nagy, A. & Rini, J. M. Simple piggyBac transposon-based mammalian cell expression system for inducible protein production. *Proc. Natl Acad. Sci. USA* **110,** 5004–5009 (2013).
84. Caro, L. N. *et al.* Rapid and facile recombinant expression of bovine rhodopsin in HEK293S GnTI(-) cells using a PiggyBac inducible system. *Methods Enzymol.* **556,** 307–330 (2015).
85. Reeves, P. J., Callewaert, N., Contreras, R. & Khorana, H. G. Structure and function in rhodopsin: high-level expression of rhodopsin with restricted and homogeneous N-glycosylation by a tetracycline-inducible N-acetylglucosaminyltransferase I-negative HEK293S stable mammalian cell line. *Proc. Natl Acad. Sci. USA* **99,** 13419–13424 (2002).
86. Wang, W. *et al.* Chromosomal transposition of PiggyBac in mouse embryonic stem cells. *Proc. Natl Acad. Sci. USA* **105,** 9290–9295 (2008).
87. Bayburt, T. H., Leitz, A. J., Xie, G., Oprian, D. D. & Sligar, S. G. Transducin activation by nanoscale lipid bilayers containing one and two rhodopsins. *J. Biol. Chem.* **282,** 14875–14881 (2007).
88. Hanson, S. M. *et al.* Differential interaction of spin-labeled arrestin with inactive and active phosphorhodopsin. *Proc. Natl Acad. Sci. USA* **103,** 4900–4905 (2006).
89. Hanson, S. M. *et al.* Structure and function of the visual arrestin oligomer. *EMBO J.* **26,** 1726–1736 (2007).
90. Fleissner, M. R., Cascio, D. & Hubbell, W. L. Structural origin of weakly ordered nitroxide motion in spin-labeled proteins. *Protein Sci.* **18,** 893–908 (2009).
91. Lietzow, M. A. & Hubbell, W. L. Motion of spin label side chains in cellular retinol-binding protein: correlation with structure and nearest-neighbor interactions in an antiparallel beta-sheet. *Biochemistry* **43,** 3137–3151 (2004).
92. Qiu, H. & Caffrey, M. The phase diagram of the monoolein/water system: metastability and equilibrium aspects. *Biomaterials* **21,** 223–234 (2000).
93. Misquitta, Y. *et al.* Rational design of lipid for membrane protein crystallization. *J. Struct. Biol.* **148,** 169–175 (2004).
94. Misquitta, L. V. *et al.* Membrane protein crystallization in lipidic mesophases with tailored bilayers. *Structure* **12,** 2113–2124 (2004).
95. ICM. Manual v. 3.8 (MolSoft LLC, 2014).
96. Coin, I. *et al.* Genetically encoded chemical probes in cells reveal the binding path of urocortin-I to CRF class B GPCR. *Cell* **155,** 1258–1269 (2013).
97. Arnautova, Y. A., Abagyan, R. A. & Totrov, M. Development of a new physics-based internal coordinate mechanics force field and its application to protein loop modeling. *Proteins* **79,** 477–498 (2011).



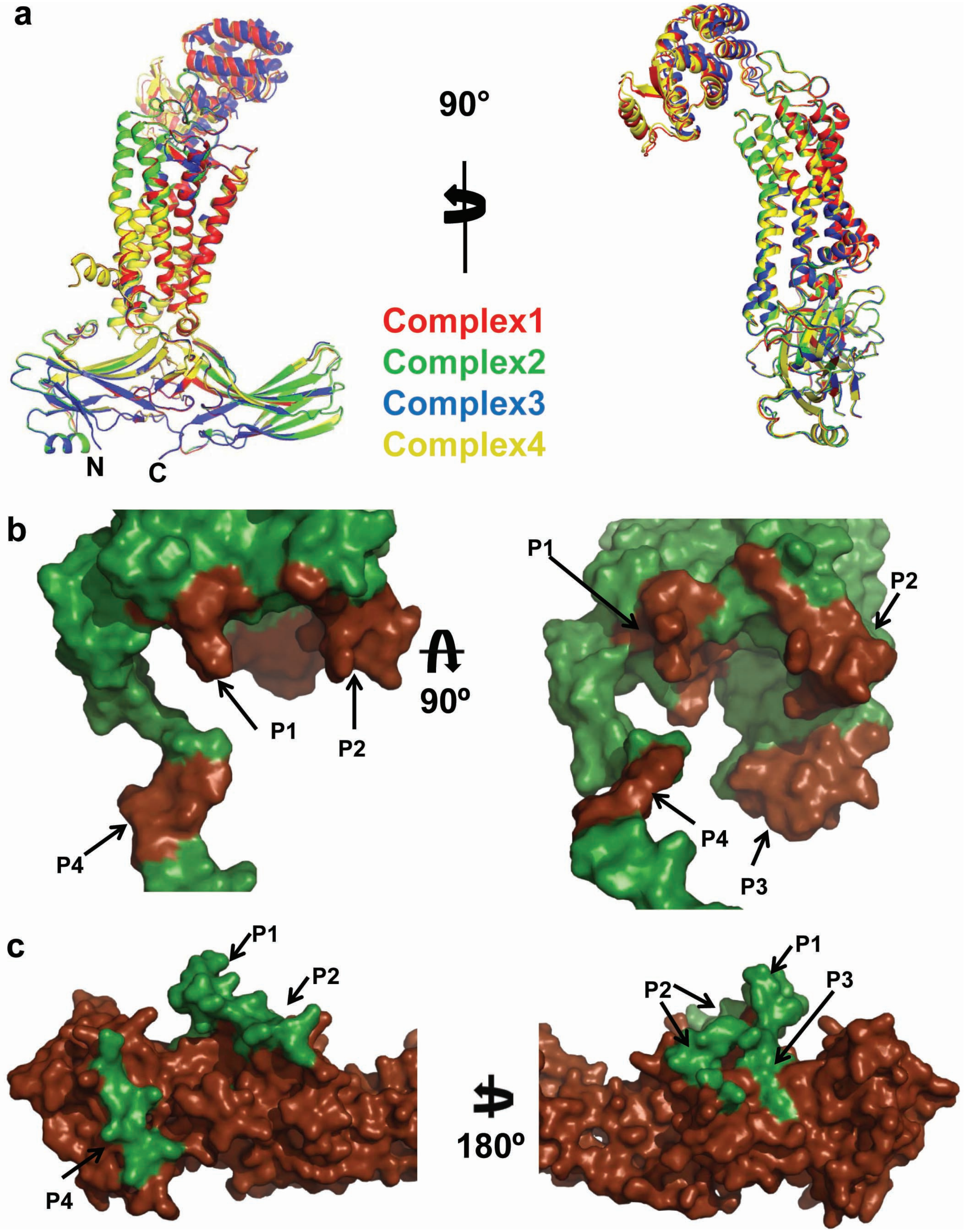
Extended Data Figure 1 | Constitutively active rhodopsin interacts with arrestin and GαCT-HA. **a**, SDS–PAGE of N-terminally MBP-tagged wild-type and mutant rhodopsin. **b**, Non-cropped versions of the pull-down assay gels shown in Fig. 1b. The interactions between mouse wild-type arrestin and human wild-type or E113$^{3.28}$Q rhodopsin are very weak. In contrast, the interaction between constitutively active rhodopsin (E113$^{3.28}$Q/M257$^{6.40}$Y) and pre-activated L374A/V375A/F376A arrestin (3A arrestin) is strong and is further increased in the presence of 10 μM all-*trans*-retinal. Input: 5% of the binding reaction. Bottom panels show the rhodopsin loading controls. **c**, Schematic representation of the AlphaScreen assay. **d**, AlphaScreen binding assay between E113$^{3.28}$Q/M257$^{6.40}$Y rhodopsin and GαCT-HA (TGGRVLEDLKSCGLF) in the presence and absence of 5 μM all-*trans*-retinal. The two left columns show the controls with 'peptide only' and 'rhodopsin only'. ($n = 3$, error bars, s.d.). **e**, Determination of the affinity of the interaction between rhodopsin E113$^{3.28}$Q/M257$^{6.40}$Y and GαCT-HA by homologous competition. His$_6$–MBP–rhodopsin mutant protein was immobilized on Ni-acceptor beads and biotinylated GαCT-HA on streptavidin donor beads. Binding between rhodopsin and arrestin brings donor and acceptor beads into close proximity, resulting in the indicated binding signal. Non-biotinylated GαCT-HA competed for the interaction with an IC$_{50}$ of ~700 nM ($n = 3$, error bars, s.d.).


a
kDa
1
2
3
250
75
50
25
His8-2×MBP-T4L-Rho-Arr
T4L-Rho-Arr
His8-2×MBP
3C protease
b
UV280 absorption (mAU)
50
40
30
20
10
0
kDa
670
158
44
17
1.35
0
1
2
3
4
5
Retention volume (ml)
c
Normalized fluorescence %
100
80
60
40
20
0
Tm = 59 °C
20
30
40
50
60
70
80
Temperature (°C)
d
50 μM
e
f
~8.0 Å

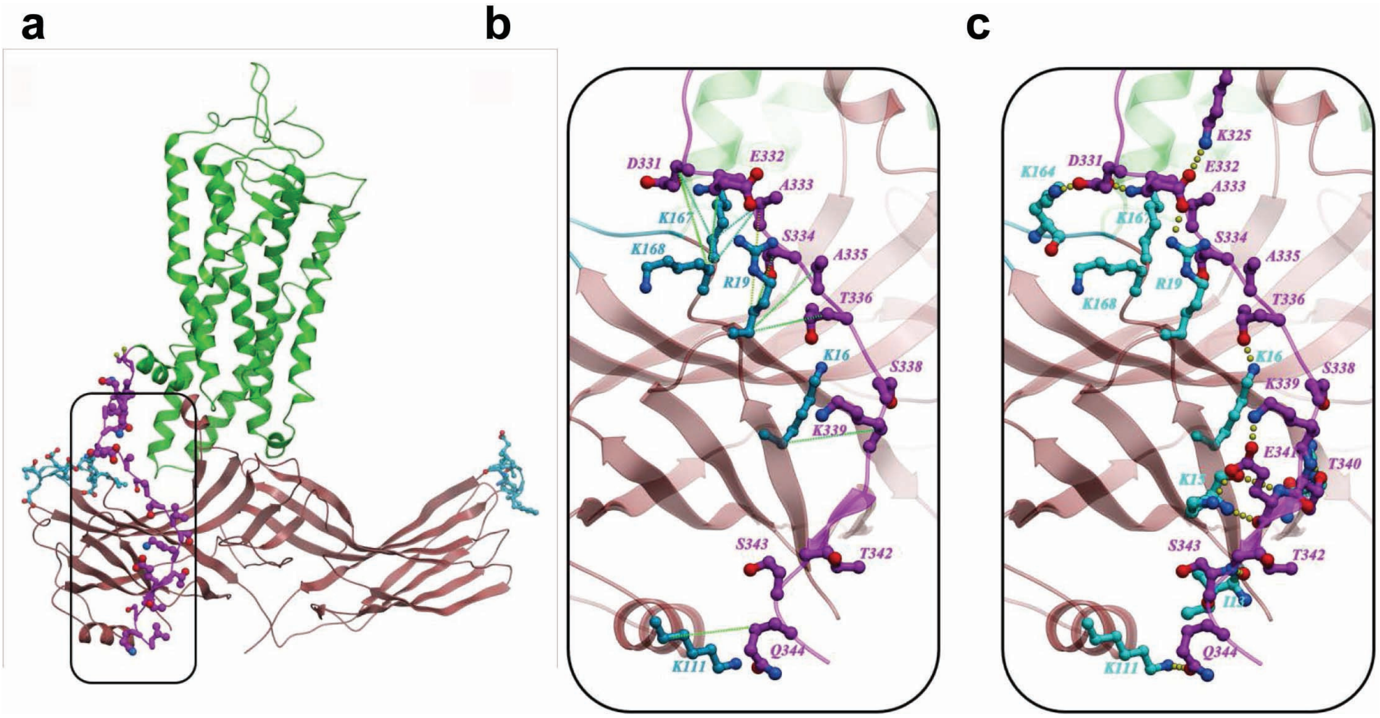
Extended Data Figure 2 | Purification and crystallization of T4L–rhodopsin–arrestin. **a**, Purification of the T4L–rhodopsin–arrestin (T4L–Rho–Arr) complex. His_8–MBP–MBP–T4L–Rho–Arr complex was first purified by amylose column chromatography (lane 1). The His_8–MBP–MBP tandem tag was then released by cleavage with 3C protease (lane 2) and removed by binding to Ni-NTA beads to recover pure T4L–rhodopsin–arrestin (T4L–Rho–Arr) protein (lane 3). **b**, Analytical gel filtration profile of the T4L–rhodopsin–arrestin complex. T4L–rhodopsin–arrestin eluted mostly as monomers with a small proportion of oligomers. The molecular weights of protein standards are indicated at the top. **c**, Thermal stability shift analysis of T4L–rhodopsin–arrestin. T4L–rhodopsin–arrestin is relatively stable with a T_m of 59 °C. **d, e**, Crystals of T4L–rhodopsin–arrestin in lipid cubic phase under bright-field illumination (**d**) and polarized light (**e**). **f**, X-ray diffraction pattern of a T4L–rhodopsin–arrestin crystal recorded at LS-CAT of APS. The green ring indicates the position of reflections at 8.0 Å resolution.



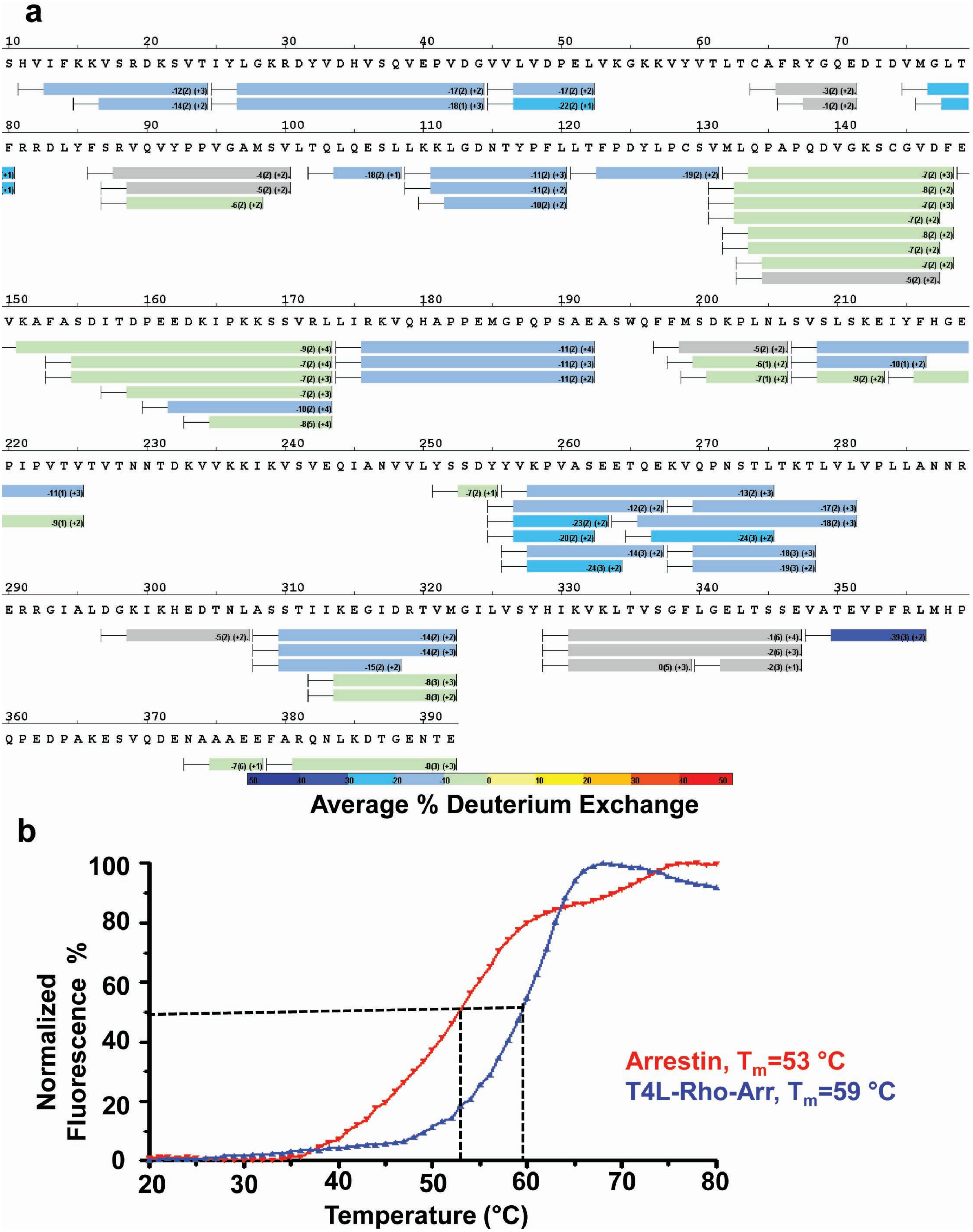
Extended Data Figure 3 | Electron density map for the overall complex and the key interfaces based on the XFEL data. **a**, A $2F_o - F_c$ electron density map contoured at 1σ of the arrestin finger loop, which forms the key interface with TM7 and helix 8. **b**, A $2F_o - F_c$ electron density map contoured at 1σ of the loop between TM5 and TM6, which forms the key interface with the β-strand following the finger loop. **c**, A 3,000 K simulated annealing omit map ($2F_o - F_c$ electron density map contoured at 1σ) calculated from the 3.8 Å/3.8 Å/3.3 Å XFEL data supports the overall arrangement of the rhodopsin–arrestin complex. In all panels, the complex structure is shown with rhodopsin coloured in green and arrestin in brown. **d**, The C-loop with a $2F_o - F_c$ composite omit map at 1σ calculated from the 3.8 Å/3.8 Å/3.3 Å truncated XFEL data. Key residues are labelled.


a
90°
Complex1
Complex2
Complex3
Complex4
N
C
b
P1
P2
P4
90º
P1
P2
P4
P3
c
P1
P2
P4
180º
P1
P2
P3

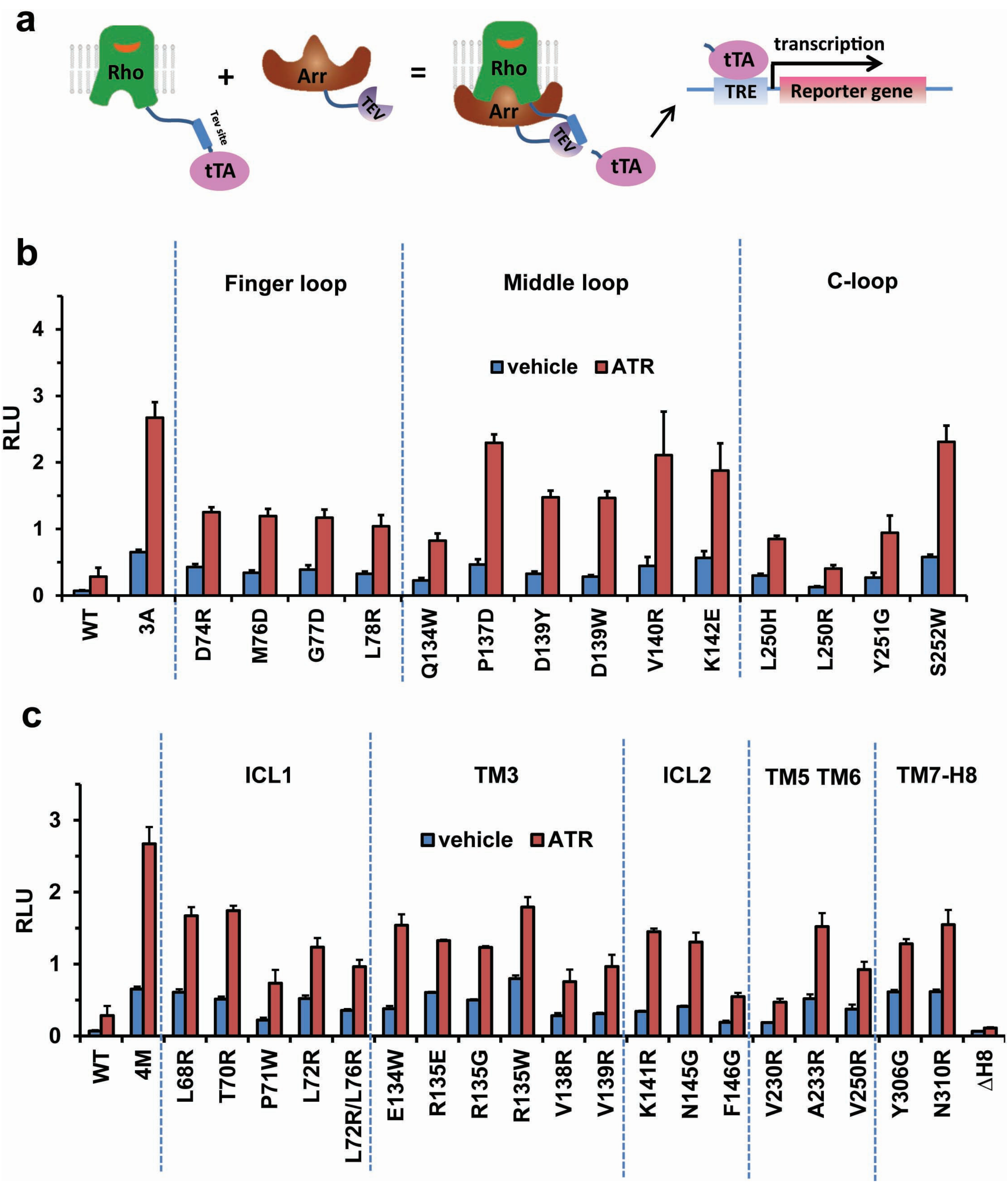
Extended Data Figure 4 | Structure similarity of the four rhodopsin–arrestin complexes in the asymmetric units and the interface between rhodopsin and arrestin. **a**, Two 90° views of the superposition of the four rhodopsin–arrestin complexes are shown as cartoon representation. The four complexes have an r.m.s.d. of less than 0.5 Å in the Cα atoms of rhodopsin and arrestin. **b**, Close-up view of arrestin-binding sites in rhodopsin. The four arrestin-binding sites (P1–P4) are highlighted in brown on the rhodopsin surface. The rhodopsin C-terminal tail/arrestin interface (P4) is based on computational modelling and disulfide cross-linking data. **c**, Rhodopsin-binding sites in arrestin. The four rhodopsin-binding sites (P1–P4) are highlighted in green on the arrestin surface.



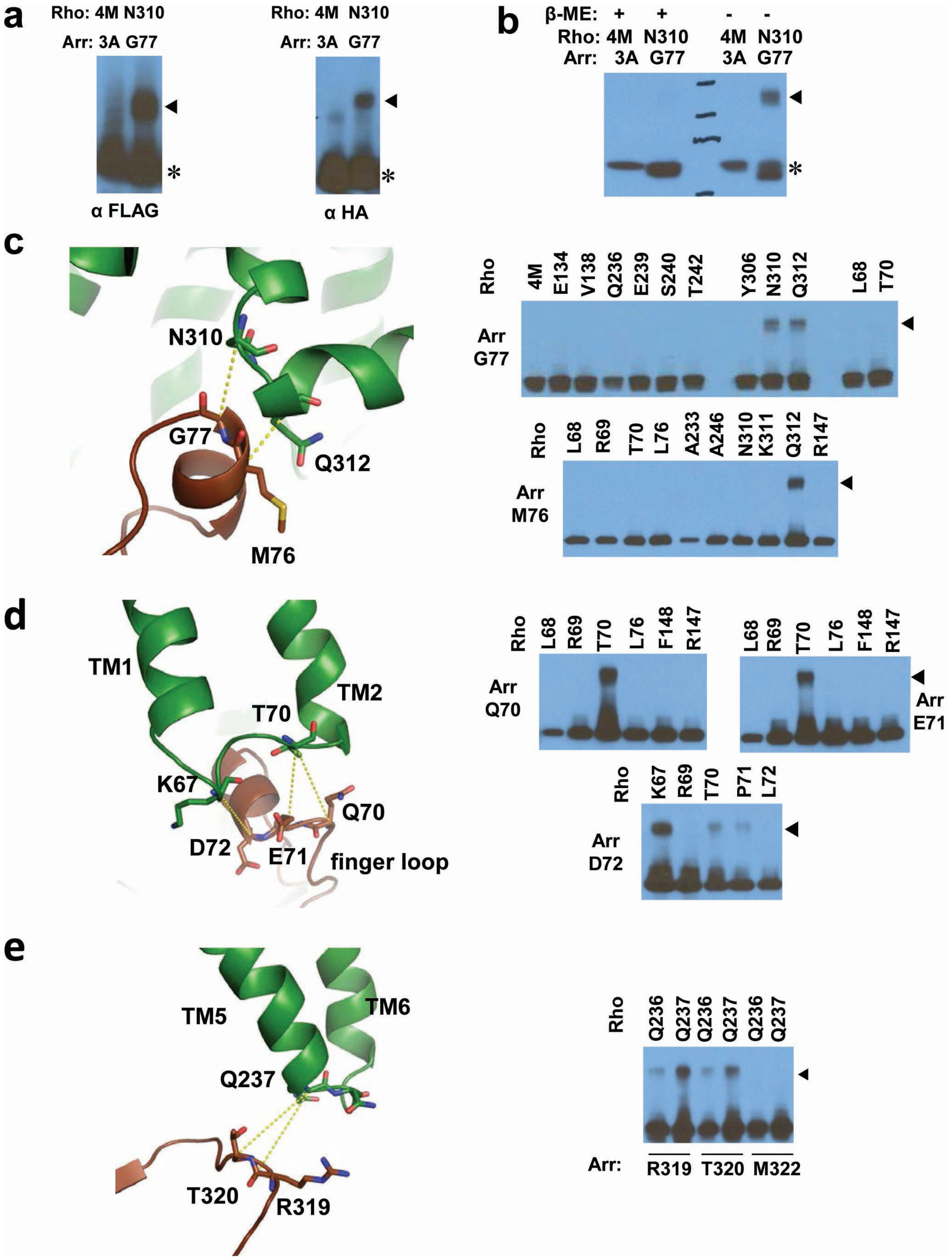
Extended Data Figure 5 | Conformational modelling of the rhodopsin–arrestin full length complex. **a**, An overview of the computational model. **b**, Predicted interactions of the rhodopsin C terminus with arrestin, showing strong to medium pairwise restraints between Cβ atoms of rhodopsin and arrestin residues identified by disulfide crosslinking. **c**, Same as in **b**, but showing predicted hydrogen bonding and ionic interactions for the C-terminal residues of rhodopsin.


a
10 20 30 40 50 60 70
S H V I F K K V S R D K S V T I Y L G K R D Y V D H V S Q V E P V D G V V L V D P E L V K G K K V Y V T L T C A F R Y G Q E D I D V M G L T
-12(2) (+3) -17(2) (+2) -17(2) (+2) -3(2) (+2)
-14(2) (+2) -18(1) (+3) -22(2) (+1) -1(2) (+2)
80 90 100 110 120 130 140
F R R D L Y F S R V Q V Y P P V G A M S V L T Q L Q E S L L K K L G D N T Y P F L L T F P D Y L P C S V M L Q P A P Q D V G K S C G V D F E
+1) -4(2) (+2) -18(2) (+1) -11(2) (+3) -19(2) (+2) -7(2) (+3)
+1) -5(2) (+2) -11(2) (+2) -8(2) (+2)
-6(2) (+2) -10(2) (+2) -7(2) (+3)
-7(2) (+2)
-8(2) (+2)
-7(2) (+2)
-7(2) (+2)
-5(2) (+2)
150 160 170 180 190 200 210
V K A F A S D I T D P E E D K I P K K S S V R L L I R K V Q H A P P E M G P Q P S A E A S W Q F F M S D K P L N L S V S L S K E I Y F H G E
-9(2) (+4) -11(2) (+4) -5(2) (+2)
-7(2) (+4) -11(2) (+3) -6(1) (+2) -10(1) (+2)
-7(2) (+3) -11(2) (+2) -7(1) (+2) -9(2) (+2)
-7(2) (+3)
-10(2) (+4)
-8(5) (+4)
220 230 240 250 260 270 280
P I P V T V T V T N N T D K V V K K I K V S V E Q I A N V V L Y S S D Y Y V K P V A S E E T Q E K V Q P N S T L T K T L V L V P L L A N N R
-11(1) (+3) -7(2) (+1) -13(2) (+3)
-12(2) (+2) -17(2) (+3)
-9(1) (+2) -23(2) (+2) -18(2) (+3)
-20(2) (+2) -24(3) (+2)
-14(3) (+2) -18(3) (+3)
-24(3) (+2) -19(3) (+2)
290 300 310 320 330 340 350
E R R G I A L D G K I K H E D T N L A S S T I I K E G I D R T V M G I L V S Y H I K V K L T V S G F L G E L T S S E V A T E V P F R L M H P
-5(2) (+2) -14(2) (+2) -1(6) (+4) -39(3) (+2)
-14(2) (+3) -2(6) (+3)
-15(2) (+2) 0(5) (+3) -2(3) (+1)
-8(3) (+3)
-8(3) (+2)
360 370 380 390
Q P E D P A K E S V Q D E N A A A E E F A R Q N L K D T G E N T E
-7(6) (+1) -8(3) (+3)
-50 -40 -30 -20 -10 0 10 20 30 40 50
Average % Deuterium Exchange
b
Normalized Fluorescence %
100
80
60
40
20
0
20 30 40 50 60 70 80
Temperature (°C)
Arrestin, T_m=53 °C
T4L-Rho-Arr, T_m=59 °C

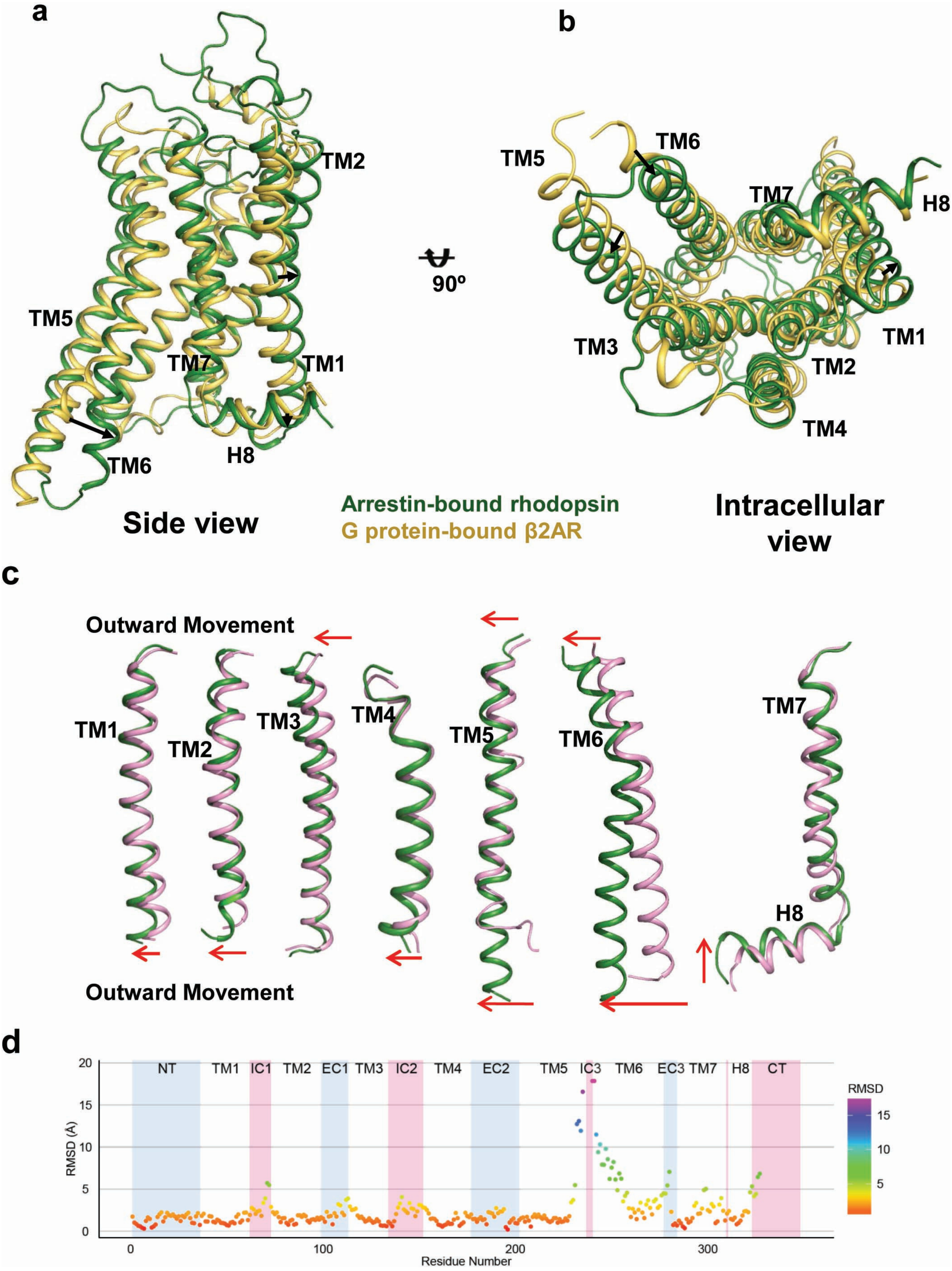
Extended Data Figure 6 | Dynamics of free 3A arrestin and rhodopsin-bound arrestin determined by HDX. **a**, HDX perturbation map between rhodopsin-bound arrestin and free arrestin, which is derived from the difference in the HDX rate between rhodopsin-bound arrestin and free arrestin. The bars below the arrestin sequence represent the peptide fragments resolved by mass spectrometry and the colours of the bars indicate the relative decrease in deuterium exchange (colour code at bottom). **b**, The thermal stability of free 3A arrestin and the rhodopsin–arrestin complex shows that the rhodopsin–arrestin complex is more stable than free 3A arrestin.



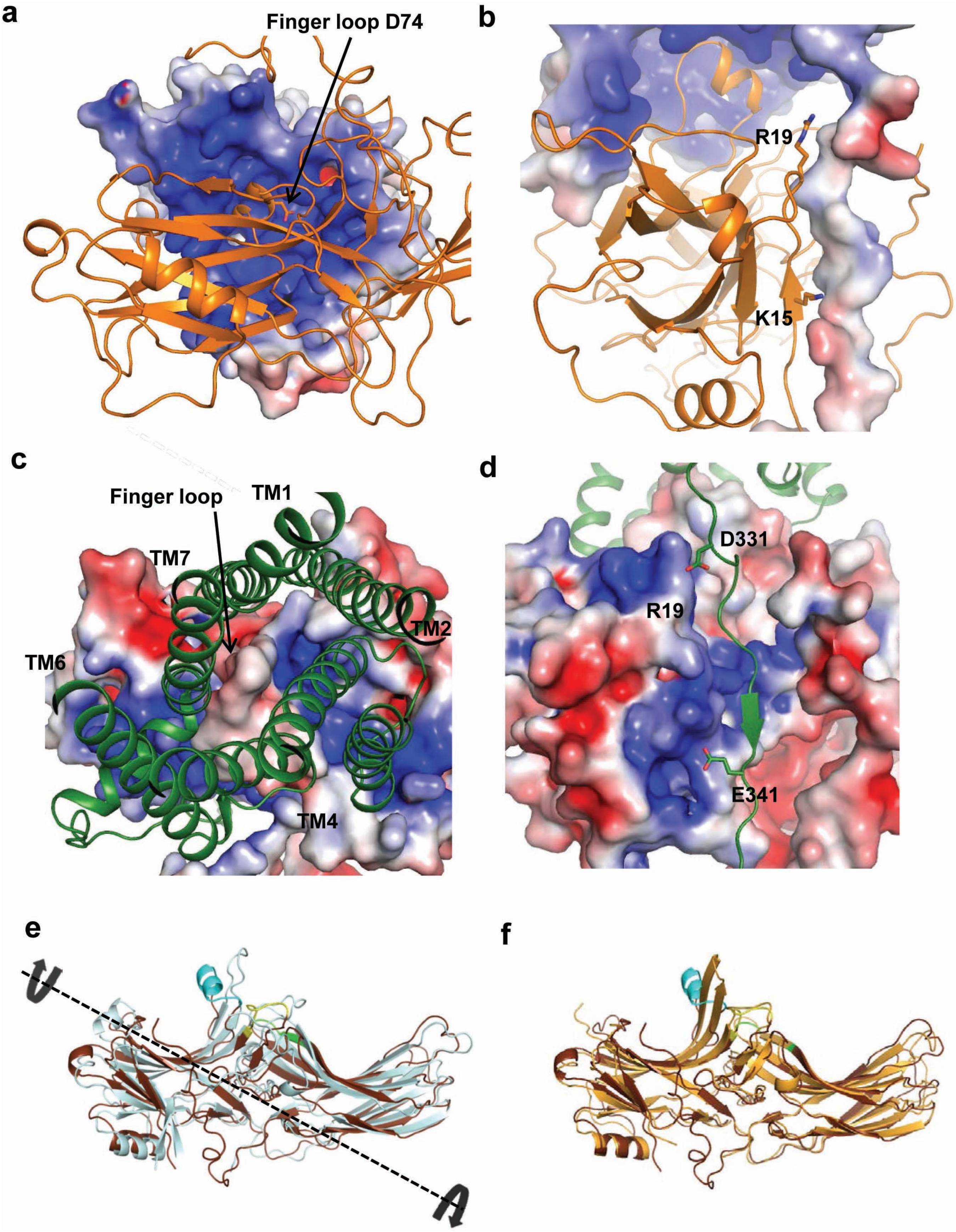
Extended Data Figure 7 | Cell-based Tango assays to validate the rhodopsin–arrestin interface. **a**, Cartoon illustration of the Tango assay for rhodopsin–arrestin interactions in cells. **b**, **c**, Mutations of key arrestin (**b**) and rhodopsin (**c**) residues that mediate the rhodopsin–arrestin interactions. Tango assay were performed in the absence or presence of 10 μM all-*trans*-retinal (ATR). ($n = 3$, error bars, s.d.).


a
Rho: 4M N310
Arr: 3A G77
α FLAG
Rho: 4M N310
Arr: 3A G77
α HA
*
b
β-ME: + + - -
Rho: 4M N310 4M N310
Arr: 3A G77 3A G77
*
c
N310
G77
Q312
M76
Rho 4M E134 V138 Q236 E239 S240 T242 Y306 N310 Q312 L68 T70
Arr G77
Rho L68 R69 T70 L76 A233 A246 N310 K311 Q312 R147
Arr M76
d
TM1
TM2
T70
K67
Q70
D72
E71
finger loop
Rho L68 R69 T70 L76 F148 R147
Arr Q70
Rho L68 R69 T70 L76 F148 R147
Arr E71
Rho K67 R69 T70 P71 L72
Arr D72
e
TM5
TM6
Q237
T320
R319
Rho Q236 Q237 Q236 Q237 Q236 Q237
Arr: R319 T320 M322

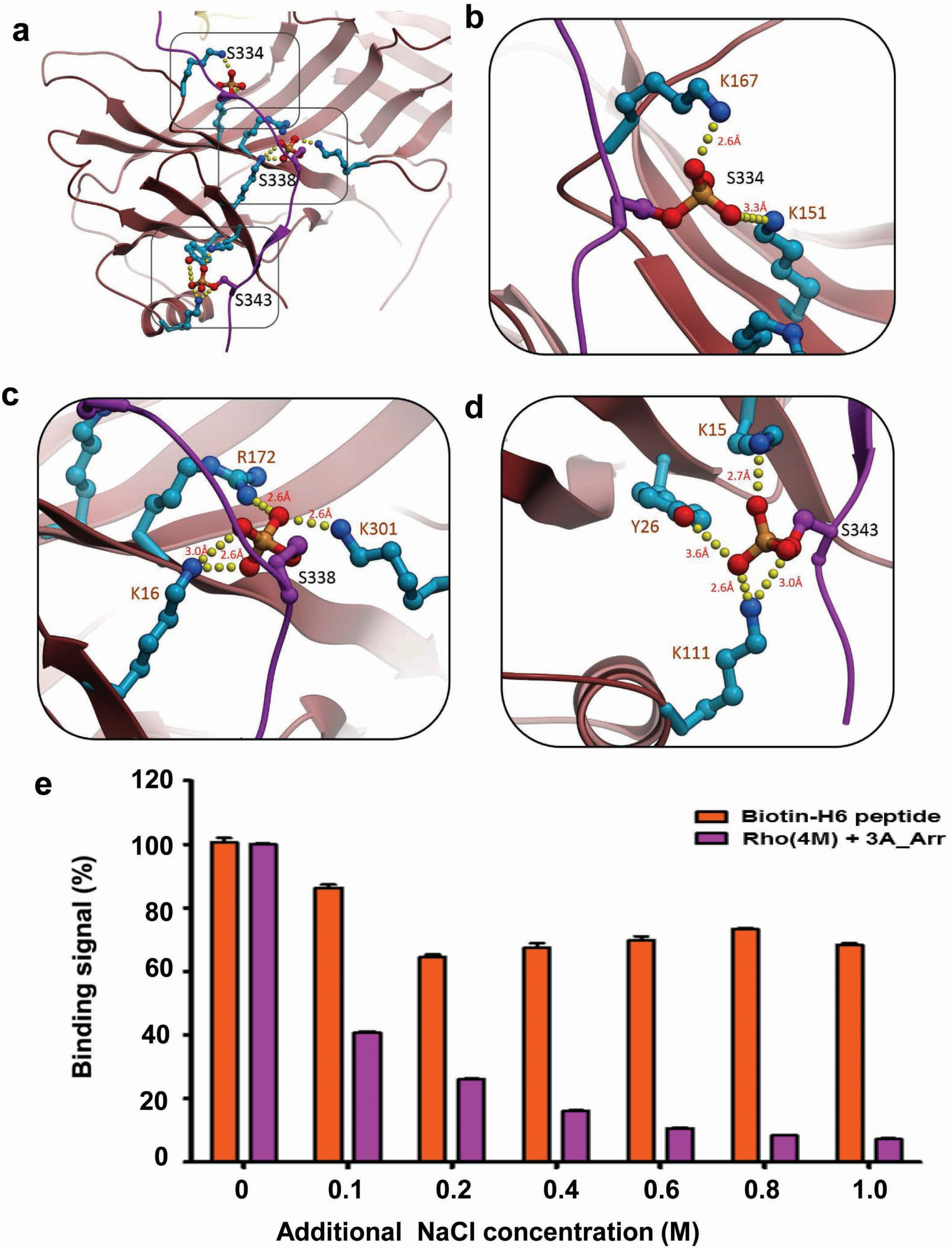
Extended Data Figure 8 | Control experiments for disulfide bond cross-linking specificity. **a**, The product of the cross-linking reaction of finger loop residue G77C with N310$^{7.57}$C of TM7 was confirmed by western blots using anti-Flag antibody (which detects arrestin–Flag fusion) and anti-HA antibody (which detects rhodopsin–HA fusion). The cross-linked products are marked with arrow heads, and free-arrestin and free-rhodopsin are indicated by asterisks. Arrestin (3A) and rhodopsin (4M) without cysteine mutations do not form cross-linked products. **b**, The cross-linked product of finger loop residue G77C with N310$^{7.57}$C of TM7 was sensitive to treatment with reducing agents, indicating the cross-linking is mediated through disulfide bond formation. **c**, A close-up view of arrestin finger loop residues M76C and G77C and their cross-linking with rhodopsin, which shows that G77C was specifically cross-linked to N310$^{7.57}$C of TM7 and Q312$^{8.49}$ of helix 8, and M76C was cross-linked to N310$^{7.57}$C of TM7 and Q312$^{8.49}$C of helix 8, but not to other residues. **d**, Structure and cross-linking of finger loop N-terminal residues Q70C, E71C, and D72C of arrestin to T70C and K67C from ICL1 of rhodopsin. **e**, Structure and cross-linking of arrestin back loop residues R319C and T320C to Q237$^{\mathrm{ICL3}}$C from TM5 of rhodopsin.


a
TM2
TM5
TM7
TM1
TM6
H8
Side view
90º
b
TM5
TM6
TM7
H8
TM3
TM1
TM2
TM4
Intracellular view
Arrestin-bound rhodopsin
G protein-bound β2AR
c
Outward Movement
TM1
TM2
TM3
TM4
TM5
TM6
TM7
H8
Outward Movement
d
NT TM1 IC1 TM2 EC1 TM3 IC2 TM4 EC2 TM5 IC3 TM6 EC3 TM7 H8 CT
RMSD (Å)
0
5
10
15
20
0
100
200
300
Residue Number
RMSD
15
10
5

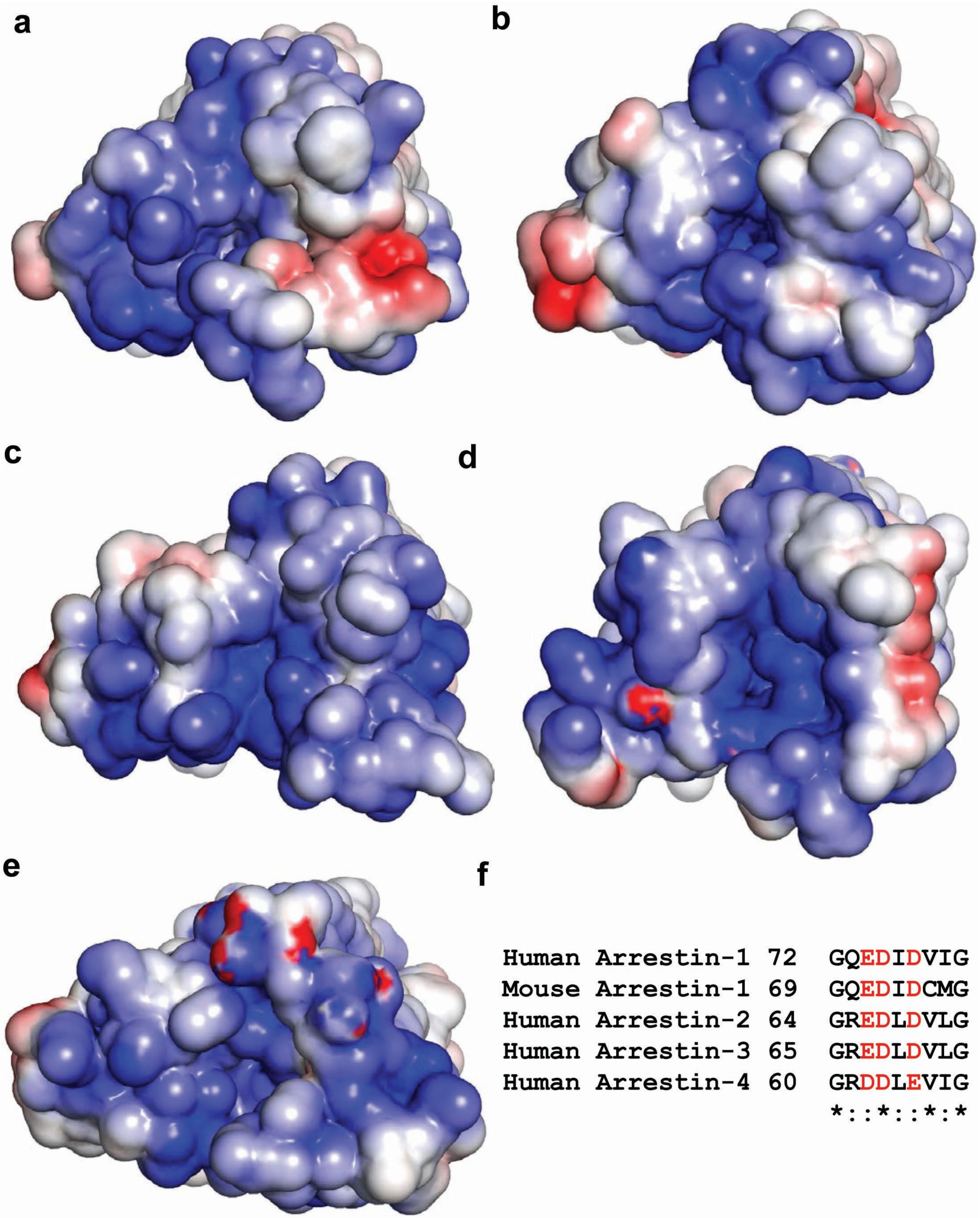
Extended Data Figure 9 | Structure comparison of the arrestin-bound rhodopsin with the β_2-adrenergic receptor in complex with G_s protein (PDB code 3SN6) and the inactive rhodopsin (PDB code 1F88). **a**, Superposition of arrestin-bound rhodopsin (green) with G_s protein-bound β_2 adrenergic receptor (light yellow). The major conformational changes are indicated by arrows. **b**, An intracellular view of a superposition of arrestin-bound rhodopsin (green) and G_s protein-bound β_2-adrenergic receptor (light yellow). **c**, Overlays of arrestin-bound rhodopsin (green) with inactive rhodopsin (pink) reveals specific conformational changes in each TM helix. The arrows indicate outward movements of TM helices. **d**, r.m.s.d. of Cα atom differences between arrestin-bound rhodopsin and inactive rhodopsin shows the large conformational changes in TM5 and TM6.


a
Finger loop D74
b
R19
K15
c
Finger loop
TM1
TM7
TM2
TM6
TM4
d
D331
R19
E341
e
f

Extended Data Figure 10 | Structure of rhodopsin-bound arrestin and its comparison with inactive and 'pre-activated' arrestin. **a**, **b**, The charge potential surface map of rhodopsin from the rhodopsin–arrestin bound complex shows that the cytoplasmic rhodopsin TM bundle surface is positively charged (blue) whereas its C-terminal tail is negatively charged (red). **c**, **d**, Charged surface of arrestin from the rhodopsin–arrestin bound complex shows that the arrestin finger loop is negatively charged (red) and its N-terminal β-strand interface is positively charged (blue). The charge distribution in rhodopsin and arrestin is complementary to each other for their interactions. **e**, Comparison of rhodopsin-bound arrestin (light blue) with inactive arrestin (brown, PDB code: 1CF1), showing an ~20° rotation between the N- and C- domains of arrestin. **f**, Comparison of rhodopsin-bound arrestin (dark brown) with pre-activated arrestin (light brown, PDB code 4J2Q), showing conformational changes in the finger loop, which adopts an α-helical conformation (cyan) in the complex. The extended finger loop conformation would protrude into the rhodopsin TM bundle and is not compatible with receptor binding. Computational model for the full rhodopsin–arrestin complex is shown in panels **b** and **d**.

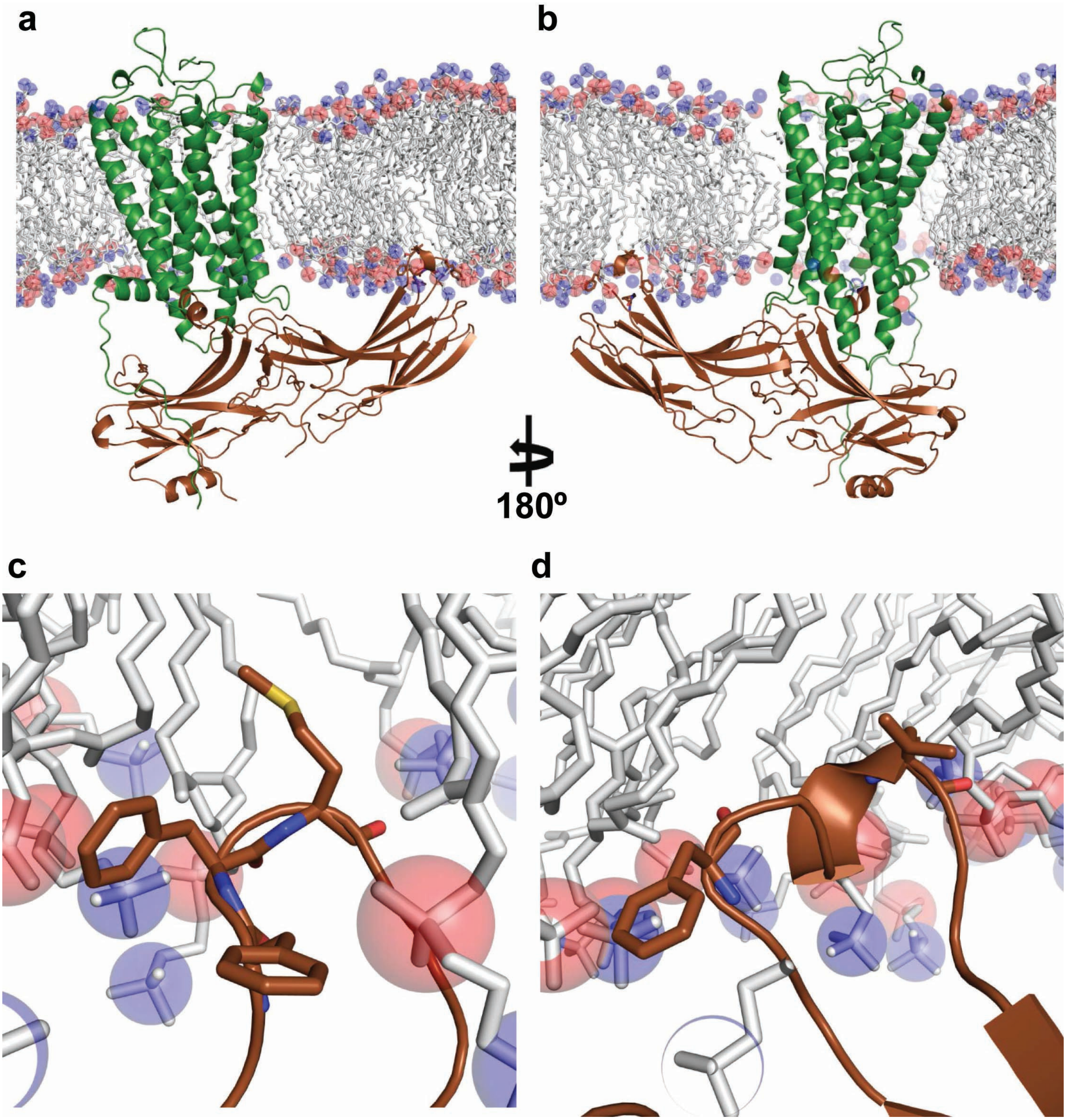

a
S334
S338
S343
b
K167
2.6Å
S334
3.3Å
K151
c
R172
2.6Å
2.6Å
K301
3.0Å
2.6Å
K16
S338
d
K15
2.7Å
Y26
S343
3.6Å
2.6Å
3.0Å
K111
e
Binding signal (%)
120
100
80
60
40
20
0
Biotin-H6 peptide
Rho(4M) + 3A_Arr
0
0.1
0.2
0.4
0.6
0.8
1.0
Additional NaCl concentration (M)

Extended Data Figure 11 | A computational model of phosphorylated rhodopsin in complex with arrestin and salt sensitivity of the rhodopsin–arrestin interaction. **a**–**d**, An overall view (**a**) and close-up views (**b**–**d**) of the computational model of the rhodopsin C-tail with phospho-serine at positions 334, 338 and 343 in complex with arrestin. **e**, The AlphaScreen control (biotin–His_6) shows much less salt sensitivity than the interaction between His-tag–rhodopsin and biotin arrestin, which is very sensitive to salt, with an IC_{50} of around 200 mM NaCl (100 mM NaCl added to 100 mM salt of the original assay buffer) ($n = 3$, error bars, s.d.).



```
Human Arrestin-1 72   GQEDIDVIG
Mouse Arrestin-1 69   GQEDIDCMG
Human Arrestin-2 64   GREDLDVLG
Human Arrestin-3 65   GREDLDVLG
Human Arrestin-4 60   GRDDLEVIG
                      *::*::*:*
```

Extended Data Figure 12 | A positive charge property is commonly found at the cytoplasmic side of GPCRs. a–e, Surface charge potential of the cytoplasmic side of selected agonist bound GPCR structures: β_1AR, PDB code 2Y02 (**a**); β_2AR, PDB code 3PDS (**b**); A_{2A} adenosine receptor, PDB code 3QAK (**c**); serotonin receptor $5HT_{1B}$, PDB code 4IAR (**d**); serotonin receptor $5HT_{2B}$, PDB code 4IB4 (**e**). Positive and negative charge potentials are shown in blue and red, respectively. **f**, Sequence alignment of the finger loop region highlighting negatively charged residues (shown in red), which are conserved in all subtypes of arrestins.



Extended Data Figure 13 | A possible role of the arrestin C-edge in lipid binding. a, **b**, The asymmetric assembly of the rhodopsin–arrestin complex in the presence of a lipid membrane bilayer, showing the C-edge of arrestin dipping into the lipid layer. **c**, **d**, A close-up view of the C-edge of arrestin in the membrane layer, where the conserved hydrophobic side chains are shown. The figure was made using the computational model for the full rhodopsin–arrestin complex.

doi:10.1038/nature14619

Magnetospherically driven optical and radio aurorae at the end of the stellar main sequence

G. Hallinan[1], S. P. Littlefair[2], G. Cotter[3], S. Bourke[1], L. K. Harding[4], J. S. Pineda[1], R. P. Butler[5], A. Golden[6], G. Basri[7], J. G. Doyle[8], M. M. Kao[1], S. V. Berdyugina[9], A. Kuznetsov[10], M. P. Rupen[11] & A. Antonova[12]

Aurorae are detected from all the magnetized planets in our Solar System, including Earth[1]. They are powered by magnetospheric current systems that lead to the precipitation of energetic electrons into the high-latitude regions of the upper atmosphere. In the case of the gas-giant planets, these aurorae include highly polarized radio emission at kilohertz and megahertz frequencies produced by the precipitating electrons[2], as well as continuum and line emission in the infrared, optical, ultraviolet and X-ray parts of the spectrum, associated with the collisional excitation and heating of the hydrogen-dominated atmosphere[3]. Here we report simultaneous radio and optical spectroscopic observations of an object at the end of the stellar main sequence, located right at the boundary between stars and brown dwarfs, from which we have detected radio and optical auroral emissions both powered by magnetospheric currents. Whereas the magnetic activity of stars like our Sun is powered by processes that occur in their lower atmospheres, these aurorae are powered by processes originating much further out in the magnetosphere of the dwarf star that couple energy into the lower atmosphere. The dissipated power is at least four orders of magnitude larger than what is produced in the Jovian magnetosphere, revealing aurorae to be a potentially ubiquitous signature of large-scale magnetospheres that can scale to luminosities far greater than those observed in our Solar System. These magnetospheric current systems may also play a part in powering some of the weather phenomena reported on brown dwarfs.

LSR J1835 + 3259 is a dwarf star of spectral type M8.5 with a bolometric luminosity $10^{-3.4}$ times that of the Sun, located at a distance of 5.67 ± 0.02 pc (ref. 4). It is positioned close to a transition in magnetic activity near the end of the main sequence on the Hertzsprung–Russell diagram, where the amount of energy produced in X-rays (indicative of the presence of a magnetically heated corona) drops by two orders of magnitude over a small range in spectral type[5]. Simultaneously, rapid rotation becomes ubiquitous, indicating a dearth of stellar-wind-assisted magnetic braking[6]. Together, these results suggest that the coolest stars and brown dwarfs possess a comparatively cool and neutral outer atmosphere relative to higher-mass (earlier spectral type) dwarf stars like our Sun. Consistent with this picture, LSR J1835 + 3259 is a rapid rotator with a period of rotation of just 2.84 h, and previous deep observations by the Chandra X-ray Observatory have failed to detect any X-ray emission associated with the presence of a magnetically heated corona[7].

LSR J1835 + 3259 has previously been observed to produce highly circularly polarized radio emission, periodically pulsed on a rotation period of 2.84 h (ref. 8). Since their initial detection as a new population of radio sources[9], similar behaviour has been observed for a number of very-low-mass stars and brown dwarfs spanning the spectral range M8 to T6.5[10,11]. In some cases, periodic variability has also been detected in broadband optical photometric bands and the Hα line[12–14]. Together, these characteristics are unlike anything observed for higher-mass main-sequence stars[15].

We pursued spectroscopic data in radio and optical bands to investigate a possible relationship between the periodic radio, broadband optical and Balmer line emission. We used the Karl G. Jansky Very Large Array (VLA) radio telescope to produce a dynamic spectrum of the periodic radio emission from LSR J1835 + 3259. Simultaneously we conducted time-resolved optical spectrophotometry using the Double Spectrograph (DBSP) on the 5.1-m Hale telescope at the Palomar Observatory. Follow-up observations, involving an additional 7 h of more sensitive time-resolved optical spectrophotometry, were carried out using the Low Resolution Imaging Spectrometer (LRIS) on the 10-m Keck telescope.

The broadband dynamic radio spectrum produced with the VLA reveals a number of distinct components periodically repeating with a 2.84-h rotation period (Fig. 1). The observed periodic features sometimes exhibit a cut-off in frequency, are 100% circularly polarized and are of very short duration relative to the rotation period, the latter implying sharp beaming. These properties are consistent with electron cyclotron maser emission produced near the electron cyclotron frequency at the source of the radio emission ($\nu_{\mathrm{MHz}} \approx 2.8 \times B_{\mathrm{Gauss}}$, where ν is frequency and B is magnetic field strength), a coherent emission process responsible for planetary auroral radio emission[2]. From the dynamic spectrum of the radio emission from LSR J1835 + 3259, we can infer magnetic field strengths in the source region of the emission of 1,550–2,850 G, close to the maximum photospheric magnetic field strengths found in late-spectral-type M7–M9 dwarfs[16]. The proximity of the radio source close to the photosphere, together with the persistent nature of the periodic radio emission, requires a current system driving a continuously propagating electron beam in the lower atmosphere of the dwarf.

The simultaneous optical spectroscopic data collected with the Hale telescope are also modulated on the 2.84-h rotational period (Fig. 1). This behaviour is confirmed with the follow-up Keck data, which has a higher signal-to-noise ratio, and for which the same periodic modulation is observed at the same amplitude. This modulation is clearly present in both spectral line emission, including the Balmer lines, and the broadband continuum optical emission of the dwarf. Most notably, the Balmer line emission and the nearby continuum vary in phase (Fig. 2), indicating a co-located region of origin.

We find that the surface feature responsible for the periodic variability in the optical spectrum can be modelled as a single component approximated as a blackbody of temperature $T \approx 2{,}200$ K, with surface coverage of <1% (Fig. 2). We attribute this blackbody-like spectrum to an optically thick region with the dominant opacity contributed by the negative hydrogen ion (H^-). H^- is the dominant source of solar

[1]California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA. [2]Department of Physics and Astronomy, University of Sheffield, Sheffield S3 7RH, UK. [3]Department of Astrophysics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK. [4]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, California 91109-0899, USA. [5]Centre for Astronomy, National University of Ireland, Galway, University Road, Galway, Republic of Ireland. [6]Department of Mathematical Sciences, Yeshiva University, New York, New York 10033, USA. [7]Astronomy Department, University of California, Campbell Hall, Berkeley, California 94720, USA. [8]Armagh Observatory, College Hill, Armagh BT61 9DG, UK. [9]Kiepenheuer Institut für Sonnenphysik, Schöneckstrasse 6, D-79104 Freiburg, Germany. [10]Institute of Solar-Terrestrial Physics, Irkutsk 664033, Russia. [11]National Radio Astronomy Observatory, PO Box O, Socorro, New Mexico 87801, USA. [12]Department of Astronomy, Faculty of Physics, St Kliment Ohridski University of Sofia, 5 James Bourchier Boulevard, 1164 Sofia, Bulgaria.

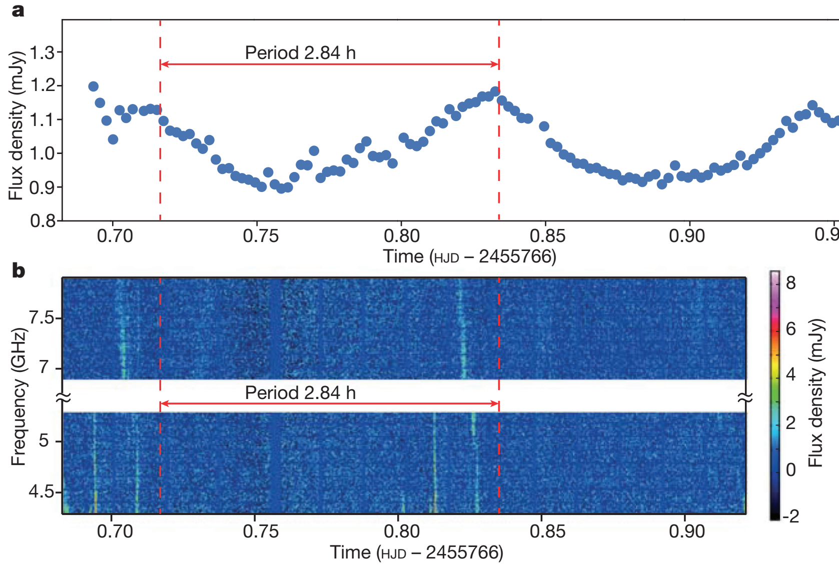


Figure 1 | Simultaneous optical and radio periodic variability of LSR J1835 + 3259. **a**, Balmer line emission extracted from spectra detected with the Hale telescope. **b**, Dynamic spectra of the right circularly polarized radio emission detected from LSR J1835 + 3259 with the VLA, with the *y* axis truncated to remove the large gap between observing bands (see Methods for details). The offset in phase of the radio features relative to the optical peak can be accounted for by the complex beaming of the radio emission. HJD, heliocentric Julian day.

optical continuum opacity, but is superseded by the opacity due to the bound–bound transitions of molecular lines, such as TiO, for cool M dwarfs, owing to the scarcity of free electrons available to form the H^- ion[17]. In the solar case, ionized metals fulfil the role of the electron donors necessary to sustain a H^- population. Since there are essentially no electron donors in a thermal gas at $T = 2{,}200$ K, another population of electron donors is required.

The periodically variable radio, Balmer line and optical continuum emission detected from LSR J1835 + 3259 can be explained by a single phenomenon, specifically a propagating electron beam striking the atmosphere, powered by auroral currents. Integrating over time and frequency, we can determine that the highly circularly polarized radio emission from LSR J1835 + 3259 contributes at least 10^{15} W of power, which requires 10^{17}–10^{19} W of power available in the electron beam for dissipation in the atmosphere, assuming an efficiency of 10^{-2}–10^{-4} for the radio emission[2]. We note that this amounts to $\sim 10^{-6}$–10^{-4} of the bolometric luminosity of the dwarf. Collisional excitation of the neutral hydrogen atmosphere by the precipitating electrons leads to subsequent radiative de-excitation, resulting in Balmer line emission with an average power of 2.5×10^{17} W, consistent with the energy budget of the precipitating electron beam. This is similar to the main Jovian auroral oval, where radio emission from electron beams contributes only 10^{-4} of the auroral power, with the bulk of the power produced in the infrared (H^{3+}; thermal), far ultraviolet (Lyman and Werner band H_2 emission) and optical (Balmer line) owing to dissipation of the electron beam energy in the atmosphere[3]. Similar ratios are also observed in Io's magnetic footprint in the Jovian atmosphere[18].

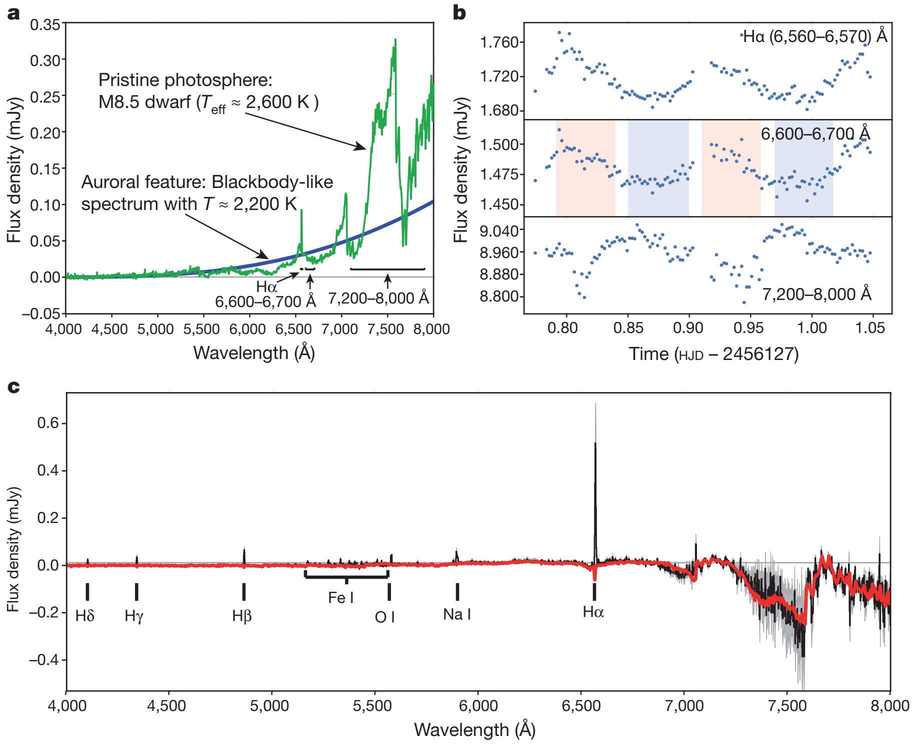


Figure 2 | Modelling the optical variability of LSR J1835 + 3259. **a**, The models we adopted for the surface brightness of the pristine photosphere (green) and auroral feature (blue). At certain wavelengths the auroral feature is brighter, whereas the pristine photosphere is brighter at other wavelengths. **b**, Lightcurves constructed from the Keck spectrophotometry for three regions of the spectrum highlighted in **a**. The lightcurve produced for Hα emission (upper panel) is tightly correlated with the lightcurve of the nearby continuum (middle panel), confirming the Balmer line emission and excess continuum emission to be approximately co-located. Meanwhile, redder wavelengths are in anti-correlation (lower panel), as expected for our model. **c**, The amplitude of variability as a function of wavelength (black) with 2σ uncertainties (shaded grey region), derived from the 'low' and 'high' states for the auroral emission, defined as the red- and blue-shaded regions respectively in **b**. The variability predicted by the model presented in **a** is shown in red (see Methods for details). We note that the line emission is not represented in the model.

We propose that the same electron beam is also causally responsible for the co-located optical broadband variability. In this model, the associated increased electron number density contributes excess free electrons, leading to an increase in the H^- population, in a process that has previously been invoked for white-light solar flares[19]. This results in an optically thick layer at a higher altitude, and thus lower temperature, than the photosphere. Despite the lower temperature, the absence of deep absorption bands in the spectrum results in a bright feature in optical bands, responsible for the broadband optical variability. However, in regions of the dwarf spectrum devoid of absorption bands, particularly towards the redder end of the spectrum, there is a reversal, with the auroral feature appearing dimmer than the surrounding photosphere. This results in some optical wavelengths displaying lightcurves that are in anti-phase to other wavelengths (Fig. 2), as previously seen in the multi-band photometry of an M9 dwarf[13].

Our observations point to a unified model involving global auroral current systems to explain the periodic radio, broadband optical and Balmer line emission detected from LSR J1835 + 3259, as well as from other low-mass stars and brown dwarfs. Extending to cooler objects, it is notable that radio emission has now been detected from a brown dwarf of spectral class T6.5 (~900 K)[11]. That brown dwarf is also notable for hosting weak Hα emission, one of only three such T dwarfs confirmed to emit Balmer line emission[20]. The similarities with LSR J1835 + 3259 suggest that the auroral mechanism operates robustly well into the regime occupied by the coolest brown dwarfs of spectral types late L and T.

It is also notable that a large degree of variability has been observed in the infrared in this spectral regime, particularly near the transition between the L and T spectral classes[21,22]. This variability explicitly requires variation in temperature or photospheric opacity across the surface of these brown dwarfs[23], which has been attributed to the spatially inhomogeneous distribution of condensate clouds in their atmospheres, effectively a manifestation of weather. This is supported by the mapping of such cloud patterns on the surface of the recently discovered Luhman 16B[24]. We speculate that the magnetospheric currents powering aurorae in brown dwarfs may also drive some of the more extreme examples of the weather phenomenon in brown dwarfs, possibly via downward propagating electron beams that modify atmospheric temperature and opacity in the same fashion as has been shown for LSR J1835 + 3259.

The nature of the electrodynamic engine powering brown-dwarf aurorae is yet to be determined. For Solar System planets, this electrodynamic engine can be (1) magnetic reconnection between the planetary magnetic field and the magnetic field carried by the solar wind (for example, Earth and Saturn)[25], (2) the departure from co-rotation with a plasma sheet residing in the planetary magnetosphere (for example, Jovian main auroral oval)[26,27], or (3) interaction between the planetary magnetic field and orbiting moons (for example, the Jupiter–Io current system)[28]. Of these, the sub-corotation of magnetospheric plasma on closed field lines, in turn powering magnetosphere–ionosphere coupling currents, has been put forward as a plausible model that can be extrapolated from the Jovian case to the brown-dwarf regime[29,30]. This model requires a continuously replenished body of plasma within the magnetosphere. This mass-loading can be achieved in multiple ways, including interaction with the interstellar medium, the sputtering of the dwarf atmosphere by auroral currents, a volcanically active orbiting planet or magnetic reconnection at the photosphere. Alternatively, considering the case of an orbiting planetary body embedded within the magnetosphere of LSR J1835 + 3259, a simple scaling from the Jupiter–Io system indicates that an Earth-sized planet (magnetized or unmagnetized) orbiting within 20 radii (<30 h orbital period) of LSR J1835 + 3259 will generate a current sufficient to power the observed aurorae. However, we note that the observed rotational modulation of the radio emission would require a substantially asymmetric magnetic field configuration for LSR J1835 + 3259. Indeed, the aurorae would display modulation on both the rotational and orbital period, which may be consistent with the large degree of variability reported for the radio pulsed emission from these objects.

A possible method of resolving the nature of the electrodynamic engine lies with the strong rotational modulation of the Balmer line emission of LSR J1835 + 3259, which implies that the auroral feature is not axisymmetric relative to the rotation axis of the dwarf. This should result in a variation in the width, intensity and velocity structure of line profiles with rotation that can be used to help map the aurora (for example, auroral oval versus polar cap), analogous to Doppler imaging, which in turn will provide information about the location of the electrodynamic engine.

Our results imply that the available power for generating aurorae on brown dwarfs is dependent on magnetic dipole moment and rotation, and may be weakly coupled to other physical characteristics, such as bolometric luminosity. This accounts for the continuous presence of auroral radio emission at similar luminosity from spectral type M8 through to T6.5, despite a decrease of two orders of magnitude in bolometric luminosity over the same spectral range[8]. This suggests that aurorae may be present at detectable levels on even the faintest T and Y dwarfs and bodes well for searches for similar emission from exoplanets.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 August 2014; accepted 26 May 2015.

1. Badman, S. V. *et al.* Auroral processes at the giant planets: energy deposition, emission mechanisms, morphology and spectra. *Space Sci. Rev.* **187,** 99–179 (2015).
2. Zarka, P. Auroral radio emissions at the outer planets: observations and theories. *J. Geophys. Res.* **103,** 20159–20194 (1998).
3. Bhardwaj, A. & Gladstone, G. R. Auroral emissions of the giant planets. *Rev. Geophys.* **38,** 295–353 (2000).
4. Reid, I. N. *et al.* Meeting the cool neighbors. IV. 2MASS 1835+32, a newly discovered M8.5 dwarf within 6 parsecs of the Sun. *Astron. J.* **125,** 354–358 (2003).
5. Stelzer, B., Micela, G., Flaccomio, E., Neuhäuser, R. & Jayawardhana, R. X-ray emission of brown dwarfs: towards constraining the dependence on age, luminosity, and temperature. *Astron. Astrophys.* **448,** 293–304 (2006).
6. Mohanty, S. & Basri, G. Rotation and activity in mid-M to L field dwarfs. *Astrophys. J.* **583,** 451–472 (2003).
7. Berger, E. *et al.* Simultaneous multi-wavelength observations of magnetic activity in ultracool dwarfs. II. Mixed trends in VB 10 and LSR 1835+32 and the possible role of rotation. *Astrophys. J.* **676,** 1307–1318 (2008).
8. Hallinan, G. *et al.* Confirmation of the electron cyclotron maser instability as the dominant source of radio emission from very low mass stars and brown dwarfs. *Astrophys. J.* **684,** 644–653 (2008).
9. Berger, E. *et al.* Discovery of radio emission from the brown dwarf LP944–20. *Nature* **410,** 338–340 (2001).
10. Hallinan, G. *et al.* Periodic bursts of coherent radio emission from an ultracool dwarf. *Astrophys. J.* **663,** L25–L28 (2007).
11. Route, M. & Wolszczan, A. The Arecibo detection of the coolest radio-flaring brown dwarf. *Astrophys. J.* **747,** L22–L25 (2012).
12. Harding, L. K. *et al.* Periodic optical variability of radio-detected ultracool dwarfs. *Astrophys. J.* **779,** 101–122 (2013).
13. Littlefair, S. P. *et al.* Optical variability of the ultracool dwarf TVLM 513–46546: evidence for inhomogeneous dust clouds. *Mon. Not. R. Astron. Soc.* **391,** L88–L92 (2008).
14. Berger, E. *et al.* Simultaneous multiwavelength observations of magnetic activity in ultracool dwarfs. I. The complex behavior of the M8.5 dwarf TVLM 513–46546. *Astrophys. J.* **673,** 1080–1087 (2008).
15. Williams, P. K. G., Cook, B. A. & Berger, E. Trends in ultracool dwarf magnetism. I. X-ray suppression and radio enhancement. *Astrophys. J.* **785,** 9–28 (2014).
16. Reiners, A. & Basri, G. A volume-limited sample of 63 M7–M9.5 dwarfs. II. Activity, magnetism, and the fade of the rotation-dominated dynamo. *Astrophys. J.* **710,** 924–935 (2010).
17. Gray, D. F. *The Observation and Analysis of Stellar Photospheres* 3rd edn, Ch. 18 (Cambridge Univ. Press, 2005).
18. Zarka, P. Plasma interactions of exoplanets with their parent star and associated radio emissions. *Planet. Space Sci.* **55,** 598–617 (2007).
19. Aboudarham, J. & Henoux, J. C. Non-thermal excitation and ionization of hydrogen in solar flares. II—Effects on the temperature minimum region energy balance and white light flares. *Astron. Astrophys.* **174,** 270–274 (1987).
20. Burgasser, A. J., Kirkpatrick, J. D., Liebert, J. & Burrows, A. The spectra of T dwarfs. II. Red optical data. *Astrophys. J.* **594,** 510–524 (2003).

21. Artigau, E., Bouchard, S., Doyon, R. & Lafreniere, D. Photometric variability of the T2.5 brown dwarf SIMP J013656.5+093347: evidence for evolving weather patterns. *Astrophys. J.* **701,** 1534–1539 (2009).
22. Radigan, J. *et al.* Large-amplitude variations of an L/T transition brown dwarf: multi-wavelength observations of patchy, high-contrast cloud features. *Astrophys. J.* **750,** 105–128 (2012).
23. Robinson, T. D. & Marley, M. S. Temperature fluctuations as a source of brown dwarf variability. *Astrophys. J.* **785,** 158–164 (2014).
24. Crossfield, I. J. M. *et al.* A global cloud map of the nearest known brown dwarf. *Nature* **505,** 654–656 (2014).
25. Isbell, J., Dessler, A. J. & Waite, J. H. Jr. Magnetospheric energization by interaction between planetary spin and the solar wind. *J. Geophys. Res.* **89,** 10716–10722 (1984).
26. Hill, T. W. The Jovian auroral oval. *J. Geophys. Res.* **106,** 8101–8108 (2001).
27. Cowley, S. W. H. & Bunce, E. J. Origin of the main auroral oval in Jupiter's coupled magnetosphere-ionosphere system. *Planet. Space Sci.* **49,** 1067–1088 (2001).
28. Goldreich, P. & Lynden-Bell, D. Io, a jovian unipolar inductor. *Astrophys. J.* **156,** 59–78 (1969).
29. Schrijver, C. J. On a transition from solar-like coronae to rotation-dominated Jovian-like magnetospheres in ultracool main-sequence stars. *Astrophys. J.* **699,** L148–L152 (2009).
30. Nichols, J. D. *et al.* Origin of electron cyclotron maser induced radio emissions at ultracool dwarfs: magnetosphere-ionosphere coupling currents. *Astrophys. J.* **760,** 59–67 (2012).

Acknowledgements We are grateful to T. Readhead, S. Kulkarni and J. McMullin for working to ensure that simultaneous Palomar and VLA observations could occur. We thank the staff of the Palomar Observatory, the W.M. Keck Observatory and the National Radio Astronomy Observatory for their support of this project. The W.M. Keck Observatory is operated as a scientific partnership among the California Institute of Technology, the University of California and the National Aeronautics and Space Administration. The Observatory was made possible by the generous financial support of the W.M. Keck Foundation. The VLA is operated by the National Radio Astronomy Observatory, a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. Armagh Observatory is grant-aided by the Northern Ireland Department of Culture, Arts and Leisure. G.H. acknowledges the generous support of D. Castleman and H. Rosen. This material is based upon work supported by the National Science Foundation under grant number AST-1212226/DGE-1144469. G.C. acknowledges support from the University of Oxford and from STFC grant ST/M006190/1. We thank J. Linsky and P. Goldreich for comments on this manuscript.

Author Contributions G.H., S.B., M.P.R., A.A., A.G., A.K., M.M.K. and J.G.D. proposed, planned and conducted the radio observations. G.H. and S.B. reduced the VLA data and the dynamic spectrum was output by S.B. G.H. interpreted the dynamic radio spectra. G.H., S.P.L., G.C., R.P.B., J.S.P. and L.K.H. proposed and conducted the Keck observations. G.C. carried out the Palomar observations and reduced the publication data. S.P.L. and G.C. reduced the Keck spectroscopic data, with the final publication data delivered by S.P.L., G.H., G.C. and S.B. G.H., S.P.L. and J.S.P. developed the interpretation of the optical data. S.P.L. carried out the detailed model fitting of the Keck spectra. G.B. analysed high-resolution archival spectra and provided insight on interpretation of the optical data. S.V.B. coordinated contemporaneous spectropolarimetry with the observations presented in this paper. S.P.L. and G.H. wrote the Supplementary Information. All authors discussed the result and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.H. (gh@astro.caltech.edu).

METHODS

Radio data reduction. Data were reduced using the Common Astronomy Software Applications (CASA Release 4.1.0; http://casa.nrao.edu/) and Astronomical Image Processing System (AIPS; http://www.aips.nrao.edu/index.shtml) packages. The amplitude and phase of the data were calibrated using short observations of the quasars QSO J1850 + 284 and 3C286 that were interspersed throughout the 6-h observation of LSR J1835 + 3259. Bad data, particularly those contaminated by radio-frequency interference, were flagged. The tasks *fixvis/UVFIX* were used to shift the source to the phase centre and the tasks *clean/IMAGR* were used to image the data. The tasks *uvsub/UVSUB* were used to subtract the source models for nearby background sources from the visibility data. The real part of the complex visibilities as a function of time and frequency for each polarization were then exported from CASA and AIPS and plotted to produce the dynamic spectrum of LSR J1835 + 3259 shown in Fig. 1.

We observed LSR J1835 + 3259 using 2 × 1-GHz sub-bands spanning frequencies of 4.3–5.3 GHz and 6.9–7.9 GHz. Two full rotation periods were captured during the 5.7-h observation. We show in Fig. 1 the dynamic spectrum detected by the right circularly polarized feeds of the VLA antennas. The original data has a time and frequency resolution of 1 s and 2 MHz respectively, but is binned and smoothed to produce the data shown in Fig. 1, with the *y* axis truncated to remove the large gap between observing bands. Periodic features of 100% circularly polarized radio emission occupy ~3% of the dynamic spectrum, allowing us to infer that it is beamed from the source in an angular emission pattern that occupies a similar fraction of a 4π steradian sphere. Studies of electron cyclotron maser emission from planetary magnetospheres have revealed the emission to be beamed in a hollow cone with walls a few degrees thick and a large opening angle (typically ~70°–90°) relative to the local magnetic field[2]. This is consistent with our data and accounts for the offset in phase of the radio features relative to the optical peak. Integrating over time and frequency and inferred beaming pattern, we determine the auroral power contributed by the polarized radio emission to be ~10^{15} W.

We detect at least six distinct components in the dynamic spectrum for LSR J1835 + 3259, each of which is probably powered by a distinct local-field-aligned current. The relationship between these individual current systems within the large-scale magnetosphere will probably remain uncertain until the nature of the electrodynamic engine is established. Although a number of these components exhibit a cut-off in emission rising to higher frequencies, there are still components present all the way to the top of the band, implying that the true cut-off in emission, associated with the largest-strength magnetic fields near the surface of the dwarf, was not captured.

Hale optical data reduction. The data from the DBSP on the Hale telescope were reduced using standard techniques with the aid of the IRAF software suite. First, bias level was subtracted from the raw frames. Then pixel-to-pixel gain variations in the charge-coupled device (CCD) camera were corrected by normalizing against exposures taken with the DBSP internal broadband lamp and the illumination function of the long slit was corrected by normalizing against twilight sky exposures. Next, the {*x*, *y*} pixels of the CCD were transformed to a rectilinear solution, {wavelength, sky position}, using the DBSP internal arc lamps. Finally, the night sky emission was subtracted by fitting a fourth-order polynomial to each column along the sky direction, with the stars in the frame masked out, and cosmic rays rejected via sigma-clipping. A tramline extraction was then used to make one-dimensional spectra of the target and reference stars.

Keck optical data reduction. Data were reduced using the LOW-REDUX pipeline (http://www.ucolick.org/~xavier/LowRedux). Bias frames were constructed by median stacking five individual bias frames. Non-uniform pixel response was removed by dividing by a dome flat field produced by stacking seven individual flat fields together. Individual objects were located on the slit, and an optimal extraction routine[31] was used to extract the object spectra on each frame. Wavelength calibration was performed using fits to arc line spectra taken at the start of the night, which gave a dispersion of 2.4 Å per pixel (red) and 3.2 Å per pixel (blue), and root-mean-square values of 0.7 pixels (red) and 0.4 pixels (blue). Each object spectrum was corrected for flexure using fits to night sky lines. Flexure corrections ranged from −2 to +2 pixels. Flux calibration was performed via a high-order polynomial fit to the flux standard Feige-110, with the Balmer lines masked out.

The data were corrected for light falling outside of the spectrograph slit using the additional comparison stars observed. Since the wavelength coverage of each object differs slightly owing to the location of slits in the object mask, slit loss corrections were determined as follows. Each comparison star was divided by an average of all the frames, to remove the spectral shape. The resulting spectrum was fitted with a first-order polynomial to give a series of wavelength-dependent slit loss corrections for each frame. This polynomial was then re-binned onto the same wavelength scale as the target spectrum. Not all comparison star spectra were used to correct for slit losses in the target spectrum. Instead, a master slit loss correction was produced via a straight mean of slit loss corrections for selected comparison stars, with the quality of slit loss correction being judged by eye. Comparison stars were removed from the slit loss correction calculations because they were either extremely blue, or not well aligned on their slits in the slit mask. LSR J1835 + 3259 was slit-loss-corrected using the two reddest comparison stars.

Wavelength dependence of optical variability. The amplitude of variability as a function of wavelength (which we term the difference spectrum) was estimated as follows. A rotational phase was assigned to each spectrum using the ephemeris of LSR J1835 + 3259. We created a spectrum representing the 'high state' and the 'low state' of LSR J1835 + 3259 by averaging all spectra with phases between 0.95–1.00 and 0.00–0.35 for the high state and 0.45–0.85 for the low state. The difference spectrum is then simply the difference between the high-state and the low-state spectra.

Statistical uncertainties on the difference spectrum are negligible compared to systematic errors, which arise from imperfect correction of slit losses, sky subtraction and removal of telluric absorption. To estimate these systematic errors, we produced a difference spectrum using an independent method. In this method, lightcurves were produced from a series of 5-Å bins, and a sinusoid of fixed phase and period was fitted to the lightcurves. A difference spectrum was produced using the amplitude of the sinusoid fitted at each wavelength. The two methods yield very similar spectra, except in the range between 7,600 Å and 7,650 Å, where the spectra are affected by telluric absorption. Subtracting the two difference spectra gave an estimate of the uncertainty, which is shown in Fig. 2.

Modelling the variability. We construct a two-phase model of the optical emission from LSR J1835 + 3259, with emission from a 'pristine' photosphere *P*, and emission from a surface feature *S*. We assume that the relative contribution from these two phases varies as the dwarf rotates. If the surface feature covers a fraction $f_{\rm h}$ during the high state and a fraction $f_{\rm l}$ during the low state, then the difference between high and low states can be written as:

$$\Delta = f_{\rm h}S + (1 - f_{\rm h})P - f_{\rm l}S - (1 - f_{\rm l})P$$

which can be simplified to:

$$\Delta = (f_{\rm h} - f_{\rm l})(S - P) = \varepsilon(S - P)$$

We model the emission from the photosphere, *P*, of LSR J1835 + 3259 using an M8 template from a SDSS library of composite M-dwarf spectra[32]. To ensure the absolute surface flux of the template is correct, we scale the template so that the bolometric flux matches that of a DUSTY model atmosphere[33] with surface gravity of log(*g*) = 5.0 and an effective temperature 2,600 K. We model the emission from the surface feature, *S*, as a blackbody with temperature $T_{\rm b}$. Our models for *S* and *P* give the flux that is crossing a unit surface of the star, so to match them to our data they need to be multiplied by a factor $N = (R/d)^2$, where *R* is the radius of LSR J1835 + 3259 and *d* is the distance to LSR J1835 + 3259. Since this is a simple multiplication of the model, this factor can be combined with the parameter ε. The two free parameters of our model are therefore $T_{\rm b}$, and the scaling constant ε.

We draw samples from the posterior distributions of our parameters by a Markov-chain Monte Carlo (MCMC) procedure. Because our difference spectrum has unknown uncertainty, a nuisance parameter σ is added. The uncertainty on the difference spectrum is set to σ everywhere, except at wavelengths corresponding to emission lines, where the uncertainty is set to an arbitrary large value.

Posterior probability distributions of $T_{\rm b}$, ε and σ are estimated using an affine-invariant ensemble sampler[34]. Uninformative priors were used for all parameters, with the exception that ε was forced to be positive. The MCMC chains consist of a total of 48,000 steps of which 24,000 were discarded as burn-in, giving 560, 770 and 860 independent samples of $T_{\rm b}$, ε and σ, respectively. The posterior probability distributions of our parameters are shown in Extended Data Fig. 1. The χ^2 of the most probable model was 860, with 920 degrees of freedom, showing that the model is an excellent fit to the data. The only wavelength regions where the model fails to reproduce our data are in the emission lines. The emission lines are probably caused by collisional excitation of the neutral atmosphere by the precipitating electrons, leading to subsequent radiative de-excitation; a process not captured by our simple two-phase model.

The best-fitting parameters are $T_{\rm b} = 2{,}180 \pm 10$ K, $\varepsilon = (1.64 \pm 0.02) \times 10^{-21}$ and $\sigma = 0.018 \pm 0.0004$ mJy. Using a model-based estimate for the radius and the measured distance for LSR J1835 + 3259 to correct ε for the $(R/d)^2$ factor, we find that the difference in covering fractions between the high state and low state is between 0.5% and 1%. These error bars do not take into account systematic uncertainties. For example, the photospheric temperature of LSR J1835 + 3259 is not determined to 10 K; adopting a different template, of M9 spectral type, for the pristine photosphere can alter $T_{\rm b}$ by approximately 50 K. Similarly, systematic errors in the scaling factor ε are probably around five times higher than the statistical errors quoted above.

Modelling the high state. By fitting the difference spectrum we are able to constrain the auroral surface feature's spectrum with a minimum of assumptions.

Nevertheless, to give confidence in our modelling, one might wish to compare our observed high-state spectrum with the predictions of our model. This requires a few additional assumptions. The high state can be written as:

$$H = N[f_h S + (1 - f_h)P]$$

Assuming the same pristine photosphere spectrum P and surface feature spectrum S that gave the best fit to our difference spectrum, we use an identical MCMC procedure to that outlined above (including a similar nuisance parameter σ for the uncertainties) to draw posterior samples of N and f_h. We find $f_h = 0.024 \pm 0.004$, $N = (8.47 \pm 0.02) \times 10^{20}$ and $\sigma = 0.196 \pm 0.003$ mJy. The resulting fit to the high-state spectrum and residuals are shown in Extended Data Fig. 2. The constraints on N above, and the constraint on ε from fitting the difference spectrum allow us to estimate $f_h - f_l = 0.0194 \pm 0.0002$ and hence $f_l = 0.005 \pm 0.004$.

A couple of pertinent features are visible in Extended Data Fig. 2. The first is that the quality of our model is limited at blue wavelengths by the signal-to-noise ratio in the M8 template spectrum that we have adopted for the pristine photosphere. Although our nuisance parameters can account for this to some degree, this is another reason why the statistical errors quoted on parameters are probably underestimates. The second is that there are features in the residuals of similar amplitude to features in the difference spectrum. These features arise because the M8 template is not a perfect fit to the pristine photosphere. However, this does not mean that our fit to the difference spectrum is unreliable. If we label our adopted photosphere template P, and the true photosphere P', then the error in the high-state spectrum will be:

$$H - H' = N(1 - f_h)(P - P')$$

whereas the error in the difference spectrum will be:

$$\Delta - \Delta' = N(f_h - f_l)(P - P')$$

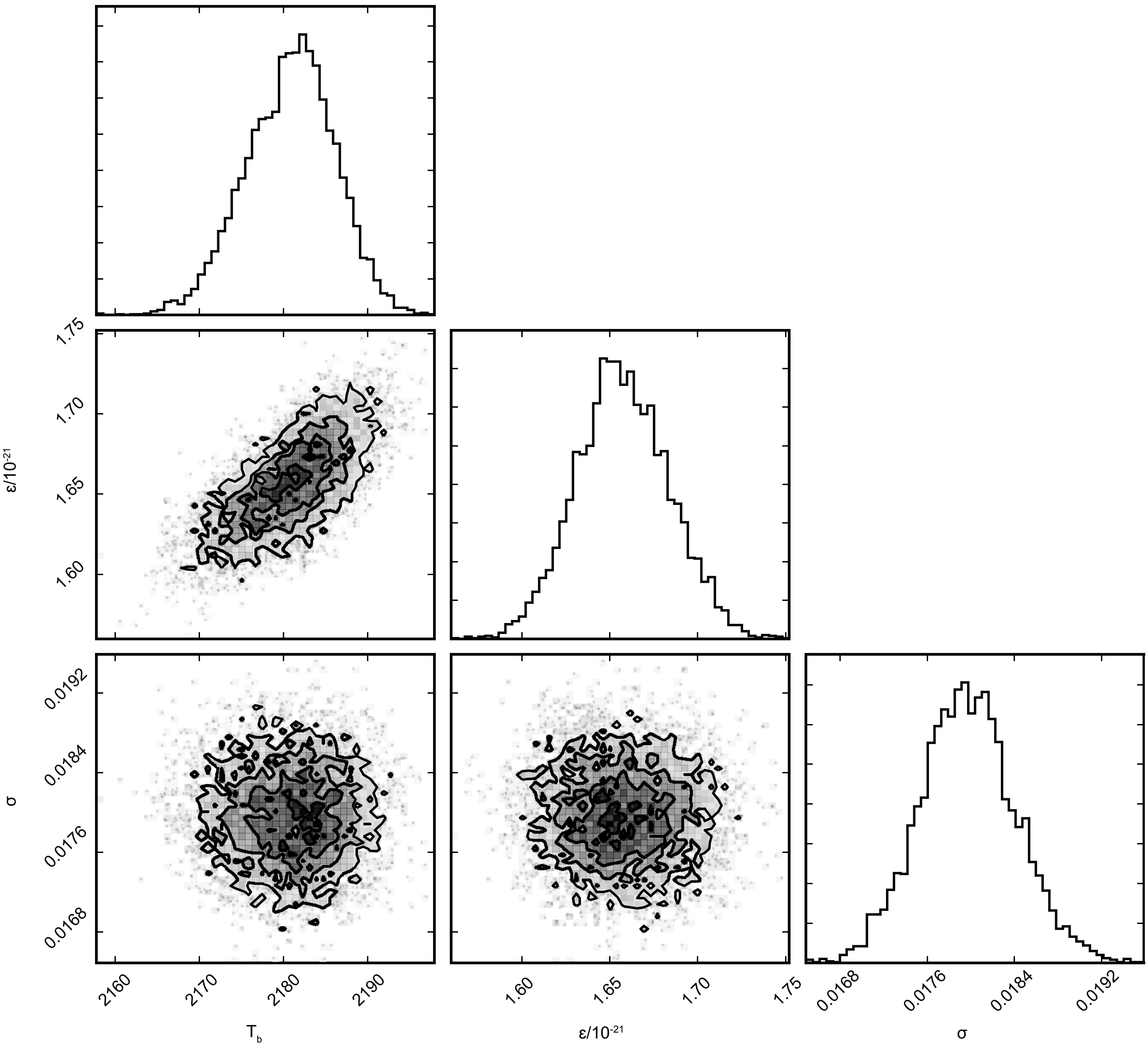
Thus, residuals in the high-state spectrum will also be present in the difference spectrum, but reduced in size by a factor of more than 50; this implies that the residuals will be smaller than the value we adopt for our nuisance parameter when fitting the difference spectrum.

An orbiting exoplanet as an electrodynamic engine. An orbiting planetary body embedded within the magnetosphere of LSR J1835 + 3259 will have motion relative to this magnetosphere and any associated frozen-in plasma. If the planet is conducting, this motion leads to the generation of an electric field across the planet that can power auroral emissions on LSR J1835 + 3259[28]. The expected power produced is proportional to the intercepted flux of magnetic energy, $P \propto vB_\perp{}^2 R_{\rm obs}{}^2$, where $B_\perp$ is the component of the magnetic field perpendicular to the planet's orbital motion, v is the planet's velocity relative to the local magnetic field and $R_{\rm obs}$ is the size of the obstacle created by the planet, the latter defined by its ionosphere or magnetosphere depending on whether the planet is magnetized or unmagnetized[17]. For example, a simple scaling from the Jupiter–Io system indicates that an unmagnetized Earth-sized planet orbiting within 20 radii (<30-h orbital period) of LSR J1835 + 3259 will generate a current sufficient to power the observed aurorae.

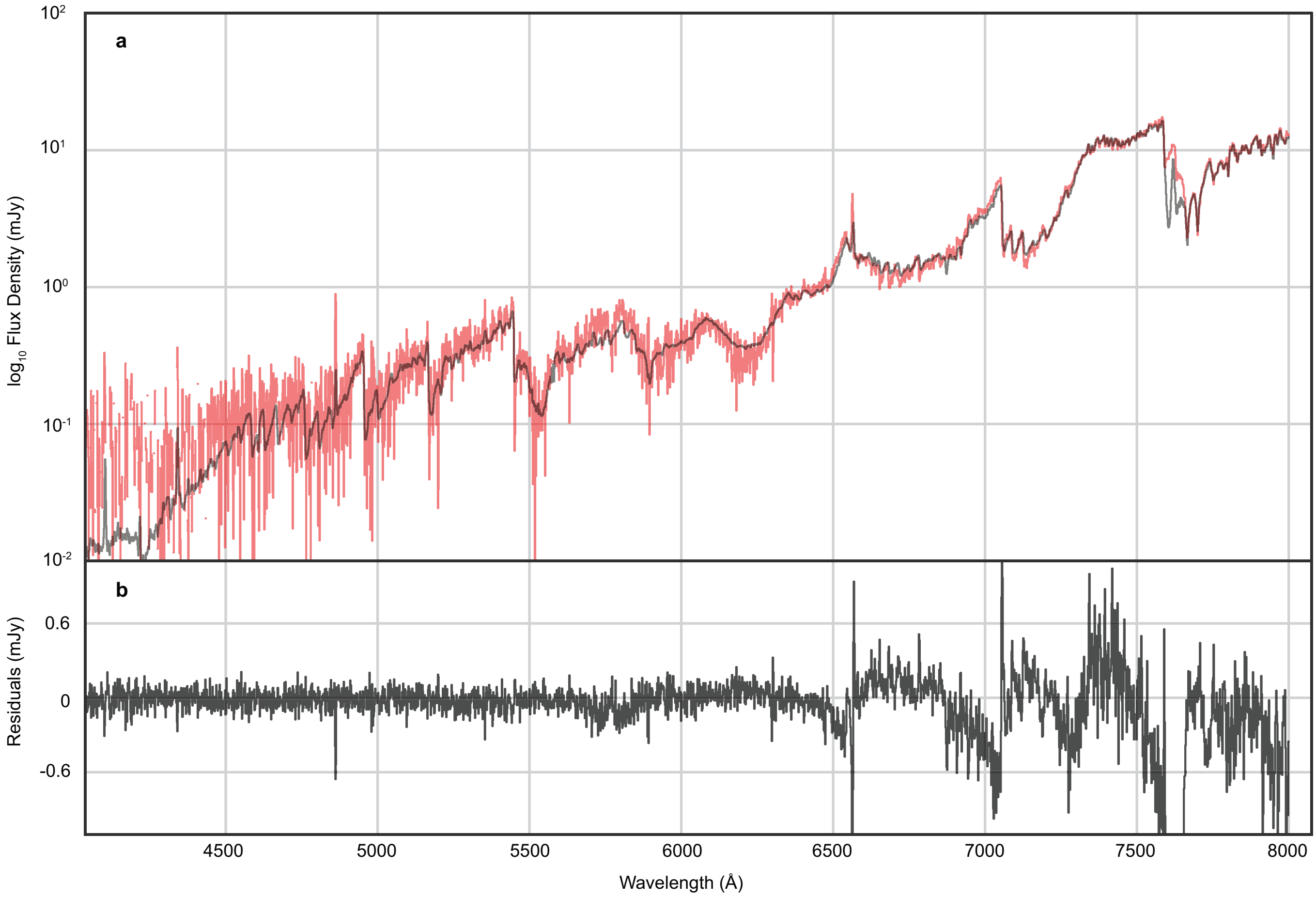
However, the resulting auroral emission is expected to be strongly modulated by the orbital period of the planet, whereas in the case of LSR J1835 + 3259, the observed periodicity of 2.84 h is consistent with rotation of the dwarf, as inferred from rotational broadening of spectral lines. Indeed, a planet orbiting with this period would be within the Roche limit of LSR J1835 + 3259 and would be torn apart by tidal forces. An alternative possibility arises if the magnetic field at the location of the planet varies substantially as LSR J1835 + 3259 rotates, which will be the case if LSR J1835 + 3259 possesses a non-axisymmetric magnetic field. For example, a tilted dipole would result in as much as fourfold variation in the current produced across the orbiting planet during each rotation of LSR J1835 + 3259. The period of the auroral emission would then be $T_{\rm aur} = T_{\rm rot}(1 + T_{\rm rot}/T_{\rm orb})$ and would approach the rotation period for large values of $T_{\rm orb}/T_{\rm rot}$. In this scenario, the aurorae would display modulation close to the rotation period as well as on the orbital period.

Code availability. The code used to model the auroral feature of LSR J1835 + 3259 is publicly available at https://github.com/StuartLittlefair/lsr1835.

31. Horne, K. Optimal spectrum extraction and other CCD reduction techniques. *Publ. Astron. Soc. Pacif.* **98,** 609–617 (1986).
32. Bochanski, J. J., West, A. A., Hawley, S. L. & Covey, K. R. Low-mass dwarf template spectra from the Sloan Digital Sky Survey. *Astron. J.* **133,** 531–544 (2007).
33. Allard, F. *et al.* The limiting effects of dust in brown dwarf model atmospheres. *Astrophys. J.* **556,** 357–372 (2001).
34. Foreman-Mackey, D., Hogg, D. W., Lang, D. & Goodman, J. emcee: the MCMC hammer. *Publ. Astron. Soc. Pacif.* **125,** 306–312 (2013).



Extended Data Figure 1 | Posterior probability distributions for two-phase model parameters. Greyscales with contours show our estimates of the joint posterior probability distributions for all combinations of parameters, while marginal posterior distributions are shown as histograms.



Extended Data Figure 2 | The high state spectrum of LSR J1835 + 3259. a, The high state spectrum of LSR J1835 + 3259 (black) is shown along with the model that best fits the high state spectrum (red). **b**, The residuals between the model and the fit.

doi:10.1038/nature14652

Real-time observation of interfering crystal electrons in high-harmonic generation

M. Hohenleutner[1]*, F. Langer[1]*, O. Schubert[1]*, M. Knorr[1], U. Huttner[2]*, S. W. Koch[2], M. Kira[2] & R. Huber[1]

Acceleration and collision of particles has been a key strategy for exploring the texture of matter. Strong light waves can control and recollide electronic wavepackets, generating high-harmonic radiation that encodes the structure and dynamics of atoms and molecules and lays the foundations of attosecond science[1–3]. The recent discovery of high-harmonic generation in bulk solids[4–6] combines the idea of ultrafast acceleration with complex condensed matter systems, and provides hope for compact solid-state attosecond sources[6–8] and electronics at optical frequencies[3,5,9,10]. Yet the underlying quantum motion has not so far been observable in real time. Here we study high-harmonic generation in a bulk solid directly in the time domain, and reveal a new kind of strong-field excitation in the crystal. Unlike established atomic sources[1–3,9,11], our solid emits high-harmonic radiation as a sequence of subcycle bursts that coincide temporally with the field crests of one polarity of the driving terahertz waveform. We show that these features are characteristic of a non-perturbative quantum interference process that involves electrons from multiple valence bands. These results identify key mechanisms for future solid-state attosecond sources and next-generation light-wave electronics. The new quantum interference process justifies the hope for all-optical band-structure reconstruction and lays the foundation for possible quantum logic operations at optical clock rates.

Ultrafast time resolution in the few-femtosecond or attosecond regime has provided systematic insight into quantum control of individual atoms[12], molecules[13], and solids[14]. A spectacular example has been to utilize the carrier wave of strong light pulses to control subcycle electron motion in atoms and molecules and follow the wavepacket dynamics directly via the temporal structure of high-order harmonic (HH) emission[1–3,11,15]. Quantum theories[16] suggest, for example, that maximum HH emission occurs at a distinct delay after the crest of the driving field, reflecting the time needed to accelerate electrons in the continuum[16,17]. Subcycle resolution has also been used to unravel novel interference phenomena in molecules[15,18,19].

In comparison, subcycle control of electrons in solids is still in its infancy, despite its promise of novel quantum physics[4–7,10,20–23] and applications in all-optical band-structure reconstruction[22,23], light-wave-driven electronics[3,5,9,10,24] or attosecond science[6–8]. Only recently high-harmonic generation (HHG) has been extended to bulk solids, setting bandwidth records in the terahertz-to-ultraviolet spectral window[4,5]. An intriguing interplay of coherent interband polarization and intraband electron acceleration in the regime of dynamical Bloch oscillations has been suggested to underlie HHG in bulk crystals[5,7,20–23]; such a process can explain, for example, the observed linear scaling of the HH cut-off frequency with the driving peak field[4,5], in contrast to a quadratic behaviour found in atoms and molecules. A detailed understanding of the microscopic electron motion as well as all envisaged applications depends critically on direct access to the temporal structure of HHs from bulk crystals[6,23,25], which has been elusive.

Here we resolve the temporal fine structure of terahertz-driven phase-locked HH pulses from a bulk semiconductor. In addition, we directly measure the HH timing with respect to the driving field on the same absolute timescale for the first time. Our data reveal that the radiation is emitted as a train of almost bandwidth-limited bursts synchronized with the maxima of the field. Differently from atoms, the bursts are emitted only during every second half-cycle. We show that these signatures originate from a new type of non-perturbative interband quantum interference involving electrons below the Fermi energy.

Multi-octave spanning HH pulses (Extended Data Fig. 1) are generated by focusing intense phase-stable multi-terahertz transients centred at a frequency of $\nu_{\mathrm{THz}} = 33\,\mathrm{THz}$ (Fig. 1a, black waveform) into a single crystal of the semiconductor gallium selenide (GaSe). To analyse the HH pulses with subcycle resolution, we introduce a novel combination of cross-correlation frequency-resolved optical gating (XFROG) and electro-optic sampling (Fig. 1a). The generated HHs and the terahertz driving field (red waveform) are superimposed with a delayed 8-fs near-infrared gate (blue waveform) and focused into a 10-μm-thick BBO (β-barium borate) crystal. Nonlinear frequency mixing simultaneously yields sum-frequency signals encoding the temporal structure of HH pulses as well as electro-optic traces of the terahertz driving waveform (see 'Experimental setup' in Methods). In this way, the relative timing of HH emission with respect to the terahertz field is determined with an uncertainty corresponding to a fraction $T/20 = 1.5\,\mathrm{fs}$ of the oscillation period T of the driving waveform (see 'Determination of the absolute timescale' in Methods and Extended Data Fig. 2).

Figure 1 compares the terahertz pump field (Fig. 1b, black curve) with the spectrally integrated (Fig. 1b, shaded curve) and the spectrally resolved (Fig. 1c, colour map) sum-frequency signal. A double-blind XFROG algorithm (see 'Double-blind XFROG algorithm' in Methods) allows us to retrieve the actual temporal envelopes and relative phases of both the gate and the HH pulses[26] from the sum-frequency data. The consistency of this analysis is confirmed by the excellent agreement between the measured and reconstructed two-dimensional spectrograms (Fig. 1c and d) and between the intensity envelope of the gate pulse retrieved from the spectrogram and an independent second harmonic FROG measurement (Extended Data Fig. 3). The retrieved time trace of the HH intensity $I_{\mathrm{HH}}(t)$ contains spectral contributions from 50 to 315 THz (Extended Data Fig. 3). $I_{\mathrm{HH}}(t)$ consists of a train of three ultrashort bursts (Fig. 1e, shaded curve) featuring three remarkable properties, as follows. (1) The maxima of $I_{\mathrm{HH}}(t)$ and $E_{\mathrm{THz}}(t)$ coincide within $\pm 2\,\mathrm{fs} = T/15$ (vertical dashed lines). This behaviour is in contrast to ballistic electron recollision models[17] where the maximum of $I_{\mathrm{HH}}(t)$ is distinctly delayed with respect to the maximal driving field[3]. (2) Unlike in atomic HHG, $I_{\mathrm{HH}}(t)$ is suppressed by one order of magnitude for field maxima of negative polarity. (3) The duration of the unipolar HH bursts is as short as 7 fs (full-width at half-maximum of intensity), which corresponds to a single oscillation period of the fourth harmonic order. Such pulse widths are expected only if all frequency components within the smooth spectral envelope (Extended Data Fig. 3) generated during one half-cycle of the driving field are emitted almost simultaneously. This is indeed the case as can be seen in Fig. 1c and d, where all sum-frequency components peak roughly at

[1]Department of Physics, University of Regensburg, 93040 Regensburg, Germany. [2]Department of Physics, University of Marburg, 35032 Marburg, Germany.
*These authors contributed equally to this work.

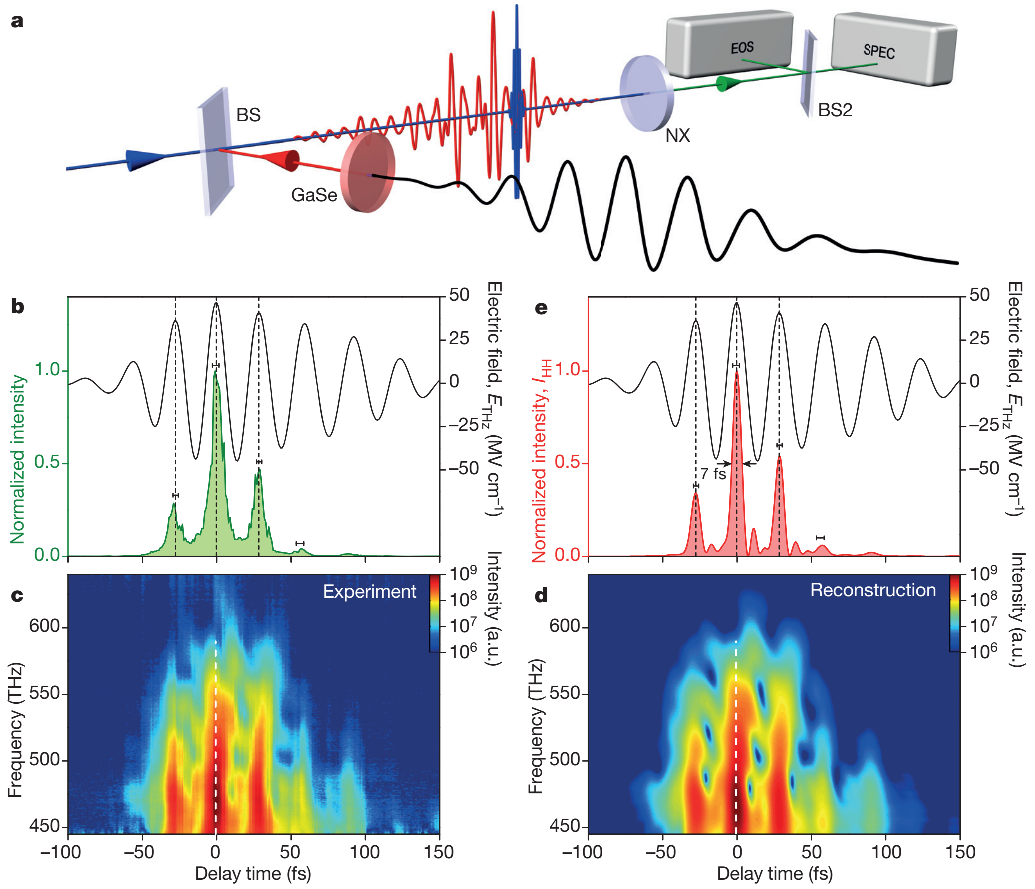


Figure 1 | Subcycle time structure of HH emission from a bulk crystalline solid. a, Experimental setup of the novel cross-correlation scheme: a multi-terahertz transient (black) is focused onto a bulk GaSe crystal (thickness 60 µm) for HH generation. The resulting waveform (red) is overlapped with a near-infrared gating pulse (blue, pulse duration 8 fs, centre wavelength 840 nm) using a beam splitter (BS) and focused into a BBO crystal (NX, thickness 10 µm) for simultaneous electro-optic interaction and sum-frequency generation. These signals (green) are split with a beamsplitter (BS2) and simultaneously recorded with a standard electro-optic sampling (EOS) setup and a spectrograph with a cooled silicon CCD detector (SPEC). **b**, Waveform of the multi-terahertz driving field featuring peak amplitudes of 47 MV cm^{-1} and a central frequency of 33 THz confirmed by electro-optic detection in a ZnTe crystal (thickness 6.5 µm, black curve). Signals obtained from sum-frequency mixing of HH and gating pulses are shown after integration over a frequency window from 490 THz to 523 THz (green curve). Dashed vertical lines highlight the local maxima of the terahertz field and error bars indicate the standard deviation of the extracted sum-frequency peak position for 12 separate measurements. **c**, **d**, Spectrograms showing the intensity of the measured sum-frequency signal for different delay times and frequencies as recorded with a Si CCD detector (**c**) and reconstructed using a double-blind XFROG algorithm (**d**), respectively. White dashed lines highlight the maximum sum-frequency intensity. **e**, Temporal shape of intensity I_{HH} (red) of the reconstructed HH pulse sequence relative to the driving multi-terahertz waveform (black). Dashed lines and error bars are the same as in **b**.

the same delay time t (vertical broken line in Fig. 1c and d), suggesting, at most, a weak spectral chirp of the HH pulses.

The observed time structure implies a quasi-instantaneous and unipolar generation mechanism. In order to identify this key ingredient, we first reproduce $I_{HH}(t)$ by a full quantum theory[5,20] (see ‘Quantum many-body model’ in Methods) including intra- and interband dynamics with two conduction and three valence bands (Extended Data Fig. 4). Our calculation reproduces the experimentally observed behaviour of $I_{HH}(t)$ in great detail (Fig. 2a, red solid curve). In particular, the emission peaks within 2 fs about the positive field crest while it is strongly suppressed for negative field extrema. In contrast, recent models accounting for only two electronic bands have consistently predicted HH emission in a bipolar fashion[7,25] and have suggested analogies with atomic HHG[22] where mostly two classes of electronic states have been considered: the ground state and the continuum of ionized states. In a solid, however, the simultaneous interaction of each electron with many atoms of the crystal lattice forms a series of electronic bands. As soon as more than two bands are included, electrons may be excited through multiple paths inducing quantum interference.

Figure 2b illustrates a minimal model for this scenario accounting for two valence bands (h_1 and h_2) and one conduction band (e_1). Excitation of an electron from band h_1 to band e_1 may either proceed by multi-photon transitions directly between two bands, $h_1 \rightarrow e_1$, or indirectly via an additional band, $h_1 \rightarrow h_2 \rightarrow e_1$. The terahertz pulse is far off either resonance, but it is sufficiently strong to generate non-perturbative excitations where electron populations change drastically on a subcycle scale. We show that such non-perturbative transitions tend to balance the respective weights of the excitation paths because the extremely strong field forces the electrons to oscillate between the non-resonantly coupled states (see ‘Interference path efficiency’ in Methods). Nonetheless, the excitation paths maintain their perturbatively assigned symmetry (see ‘Strong-field quantum interference’ in Methods), featuring an odd transition amplitude $A_o(-E_{THz}) = -A_o(E_{THz})$ with respect to the driving field for the direct excitation and an even amplitude $A_e(-E_{THz}) = A_e(E_{THz})$ for the indirect path $h_1 \rightarrow h_2 \rightarrow e_1$ (Fig. 2b, Extended Data Fig. 5). A coherent superposition of both yields a total amplitude of $A_e(|E_{THz}|) + (-) A_o(|E_{THz}|)$ for positive (negative) E_{THz}, respectively (see ‘Perturbative versus non-perturbative quantum interference’ in Methods). Hence, the sign of the field controls the total outcome of HH transitions. Note that the transition $h_1 \rightarrow h_2$ connecting bands below the Fermi level is initially Pauli-blocked but strong excitation can significantly empty h_2, for example, via the transition $h_2 \rightarrow e_1$, clearing the path $h_1 \rightarrow h_2 \rightarrow e_1$.

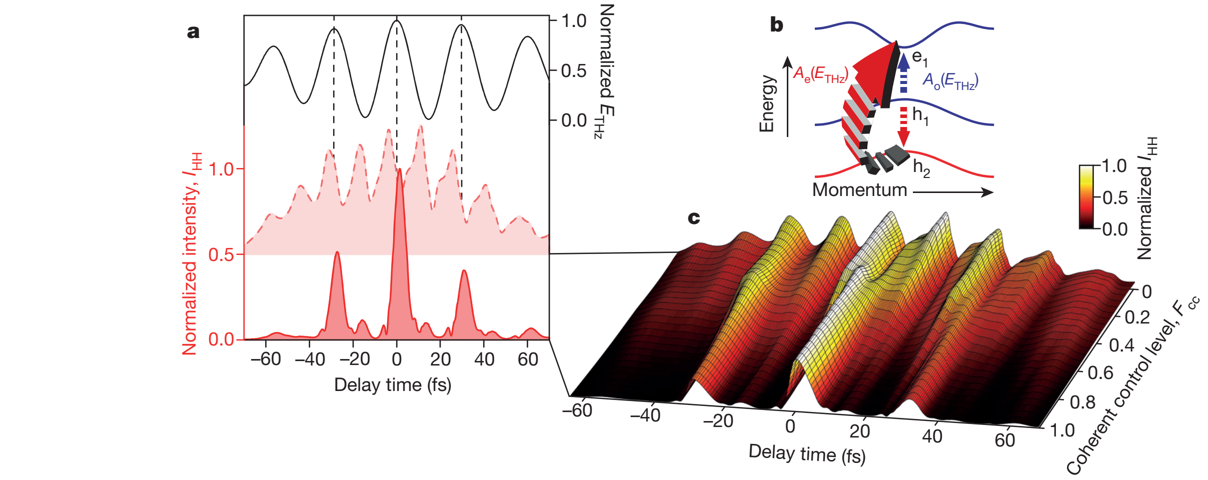


Figure 2 | Non-perturbative quantum interference in HH emission. **a**, Driving terahertz field (black curve) and calculated intensity envelope of HH emission as a function of time for $F_{cc} = 0$ (broken red curve, magnified by a factor of 25) and $F_{cc} = 1$ (solid red curve). Dashed vertical lines highlight the local maxima of the terahertz field. **b**, Simplified three-band schematic of different ionization pathways from valence band h_1 to conduction band e_1. The direct transition amplitude $A_o(E_{THz})$ is an odd function of the driving field, $A_o(-E_{THz}) = -A_o(E_{THz})$. The amplitude $A_e(E_{THz})$ of the indirect path ($h_1 \rightarrow h_2 \rightarrow e_1$) is the product of two odd functions, resulting in even symmetry $A_e(-E_{THz}) = A_e(E_{THz})$. The indirect path features a transition ($h_1 \rightarrow h_2$) that is initially blocked by Pauli exclusion and only opens under strong-field excitation. **c**, HH intensity envelopes computed within the five-band model as a function of delay time and the coherent control factor F_{cc} regulating coherent transitions between occupied valence bands. Bright colours mark strong emission, dark colours mark weaker emission (see key). All time traces are normalized separately.

We test the viability of this concept by a systematic switch-off analysis using our five-band computation that includes all relevant transitions. By artificially multiplying the dipole moment $d_{h_1h_2}$ between the hole bands h_1 and h_2 as well as all other similar terms with a coherent control factor F_{cc}, we eliminate the indirect paths needed for non-perturbative quantum interference. Figure 2a compares the intensity envelope, $I_{HH}(t)$, with (red solid line, $F_{cc} = 1$) and without (red dashed line, $F_{cc} = 0$) the indirect paths. Switching off the quantum interference, that is, considering only direct transitions ($F_{cc} = 0$), produces bursts at positive and negative crests of the field. Interestingly, the bursts become delayed by roughly $T/4$ with respect to the field extrema, which is consistent with a delay expected in an atomic recollision model[3,17]. However, opening the indirect paths ($F_{cc} = 1$) synchronizes the emission with the driving field. More specifically, the interband coherence is driven such that the quantum interference and the resulting HH emission are strongest during the presence of the electric field. This process maximizes (suppresses) the emission with the positive (negative) crests of the field. The destructive interference is not perfect, leaving small HH remnants at negative peak fields.

Figure 2c shows computed normalized $I_{HH}(t)$ traces as a function of F_{cc} (unscaled representation in Extended Data Fig. 6). By gradually suppressing the coherent-control paths, emission appears as delayed bursts after each field maximum and minimum. Nevertheless, the transition is not smooth, but contains non-trivial oscillations and bifurcations. These features are caused by terahertz-induced band mixing, which modulates electronic populations and $I_{HH}(t)$ and underpins the non-perturbative character of the interband excitations.

Under the extremely non-resonant conditions of our experiment ($h\nu_{THz} < E_g/14$ with $E_g = 2.0$ eV being the bandgap energy of GaSe), band-to-band transitions require multi-terahertz pump photons. Non-perturbative excitations can non-resonantly drive all these transitions to exhibit large population transfer (Extended Data Fig. 5), and the related processes are robust against variations of the terahertz field strength and photon energy. Therefore, the quantum interference should be detectable for a broad range of field amplitudes E_{THz} and terahertz photon energies. In fact, the HH maxima remain synchronized with the positive peak of the driving field for both experimental (Fig. 3a) and theoretical (Fig. 3b) traces of $I_{HH}(t)$ when the terahertz frequency is changed between 25 and 34 THz, whereas the temporal separation of the emission bursts grows with the oscillation period of the driving field. Both measured (Fig. 3c) and computed (Fig. 3d) traces of $I_{HH}(t)$ also remain unipolar and quasi-instantaneous when the terahertz field amplitude E_{THz} is changed. The contrast is even enhanced for higher field strengths.

Recent studies have demonstrated that strong terahertz fields, needed to create HH emission, can also coherently accelerate electrons throughout the Brillouin zone before scattering occurs[4–6,21,27]. Owing to resulting dynamical Bloch oscillations, electrons may undergo one or more Bragg reflections within one half-cycle of the driving field, emitting high-frequency radiation at the quasi-instantaneous Bloch frequency ν_B. Since ν_B is proportional to E_{THz} (see ref. 28), the frequency of the Bloch-related contribution to HHG should trace the temporal profile of the driving field. Our XFROG algorithm allows us to retrieve the temporal phase $\phi_{HH}(t)$ of the HH pulse train (Extended Data Fig. 7), from which we obtain its instantaneous frequency $\nu_i(t) = (2\pi)^{-1}\partial\phi_{HH}/\partial t$, weighted by the spectral amplitude within our detection bandwidth (Fig. 3e). All time traces of $\nu_i(t)$ measured for different terahertz amplitudes follow a universal double-chirp pattern, which is a fingerprint of dynamical Bloch oscillations: after a monotonic increase during the rising slope of $E_{THz}(t)$, ν_i peaks approximately at the maximum of the applied field and decreases again following the abating driving field. With increasing amplitude, the instantaneous frequency in a single HH pulse blue-shifts globally while its maximum broadens, develops shoulders and finally morphs into a non-monotonic pattern for the highest field strengths. Our quantum theory reproduces even these non-trivial features well (Fig. 3f).

The combination of non-perturbative quantum interference and dynamical Bloch oscillations may be systematically harnessed for ultrashort pulse shaping. By varying the terahertz carrier frequency (Fig. 3a, b) and the carrier-envelope phase (CEP) of the driving waveform (Extended Data Fig. 8), the global shape of the HH pulse sequence can be tailored, whereas the frequency modulation within individual bursts is reproducibly set by the terahertz amplitude. Almost bandwidth-limited pulses may be generated—especially if the phase-flattening effect for high peak fields is exploited (Fig. 3e, f). We expect that in our experiment, suitable high-pass filtering of the HH pulses may allow for pulse durations as short as 3 fs in the infrared and visible domain (Extended Data Fig. 9). Since the principle of solid-based HHG is fully scalable to the ultraviolet, even shorter pulses may be possible in wide-gap materials.

In conclusion, the relative timing of HHG with respect to the driving field, the unipolar response, and a non-monotonic frequency modulation

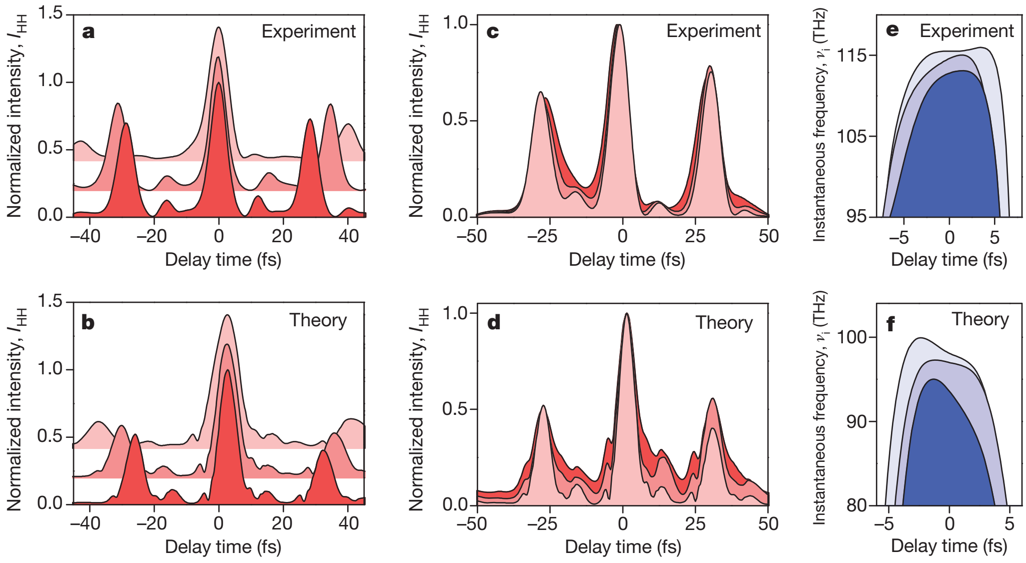


Figure 3 | Tunability and robustness of non-perturbative quantum interference. **a**, **b**, Measured (**a**) and calculated (**b**) intensity envelopes I_{HH} of emitted HH for driving fields featuring central frequencies of 25, 30 and 34 THz, respectively. Darker colours represent higher central frequencies. **c**, **d**, Measured (**c**) and calculated (**d**) HH intensity envelopes for driving peak fields of 26, 31 and 44 MV cm^{-1} (experiment) and 19, 22 and 31 MV cm^{-1} (theory). Brighter colours represent higher peak fields. **e**, **f**, Instantaneous frequency ν_i of the central HH emission bursts shown in **c** and **d**, respectively. Brighter colours represent higher peak fields.

provide direct insight into the terahertz strong-field-driven motion of electrons in GaSe. We identify a non-perturbative quantum interference between interband transitions as a salient HH generation mechanism. In its most generic form, this strong-field mechanism can occur if (1) there is a closed-loop triangle system of states that are all mutually coupled by dipole transitions (Extended Data Fig. 10), and (2) the terahertz field is far below these resonances and (3) strong enough to generate non-perturbative excitations changing carrier populations substantially (see Methods and Extended Data Fig. 5). In contrast to established techniques of perturbative quantum interference between one- and two-photon transitions inducing directed charge and spin currents[29], our new concept is robust even at extremely strong fields. Thus, it may inspire new techniques for quantum-logic operations[30] based on sturdy, non-perturbative transitions between strongly coupled energy bands. Driving coherences between initially fully occupied valence bands, HHG provides all-optical access even to details of the band structure hidden below the Fermi level. Furthermore, the direct observation of lightwave-controlled electron dynamics marks the way towards a complete microscopic picture of HHG in solids, ultrafast electronics, and novel solid-state CEP-stable attosecond sources.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 January; accepted 1 June 2015.

1. Paul, P. M. *et al.* Observation of a train of attosecond pulses from high harmonic generation. *Science* **292,** 1689–1692 (2001).
2. Dudovich, N. *et al.* Measuring and controlling the birth of attosecond XUV pulses. *Nature Phys.* **2,** 781–786 (2006).
3. Krausz, F. & Stockman, M. I. Attosecond metrology: from electron capture to future signal processing. *Nature Photon.* **8,** 205–213 (2014).
4. Ghimire, S. *et al.* Observation of high-order harmonic generation in a bulk crystal. *Nature Phys.* **7,** 138–141 (2011).
5. Schubert, O. *et al.* Sub-cycle control of terahertz high-harmonic generation by dynamical Bloch oscillations. *Nature Photon.* **8,** 119–123 (2014).
6. Ghimire, S. *et al.* Strong-field and attosecond physics in solids. *J. Phys. B* **47,** 204030 (2014).
7. Higuchi, T., Stockman, M. I. & Hommelhoff, P. Strong-field perspective on high-harmonic radiation from bulk solids. *Phys. Rev. Lett.* **113,** 213901 (2014).
8. Mücke, O. D. Isolated high-order harmonics pulse from two-color-driven Bloch oscillations in bulk semiconductors. *Phys. Rev. B* **84,** 081202 (2011).
9. Goulielmakis, E. *et al.* Attosecond control and measurement: lightwave electronics. *Science* **317,** 769–775 (2007).
10. Zaks, B., Liu, R. B. & Sherwin, M. S. Experimental observation of electron–hole recollisions. *Nature* **483,** 580–583 (2012).
11. Shafir, D. *et al.* Resolving the time when an electron exits a tunnelling barrier. *Nature* **485,** 343–346 (2012).
12. Drescher, M. *et al.* Time-resolved atomic inner-shell spectroscopy. *Nature* **419,** 803–807 (2002).
13. Calegari, F. *et al.* Ultrafast electron dynamics in phenylalanine initiated by attosecond pulses. *Science* **346,** 336–339 (2014).
14. Neppl, S. *et al.* Direct observation of electron propagation and dielectric screening on the atomic length scale. *Nature* **517,** 342–346 (2015).
15. Smirnova, O. *et al.* High harmonic interferometry of multi-electron dynamics in molecules. *Nature* **460,** 972–977 (2009).
16. Salières, P. *et al.* Feynman's path-integral approach for intense-laser-atom interactions. *Science* **292,** 902–905 (2001).
17. Corkum, P. B. Plasma perspective on strong-field multiphoton ionization. *Phys. Rev. Lett.* **71,** 1994–1997 (1993).
18. Shafir, D., Mairesse, Y., Villeneuve, D. M., Corkum, P. B. & Dudovich, N. Atomic wavefunctions probed through strong-field light-matter interaction. *Nature Phys.* **5,** 412–416 (2009).
19. Kanai, T., Minemoto, S. & Sakai, H. Quantum interference during high-order harmonic generation from aligned molecules. *Nature* **435,** 470–474 (2005).
20. Golde, D., Meier, T. & Koch, S. W. High harmonics generated in semiconductor nanostructures by the coupled dynamics of optical inter- and intraband excitations. *Phys. Rev. B* **77,** 075330 (2008).
21. Földi, P., Benedict, M. G. & Yakovlev, V. S. The effect of dynamical Bloch oscillations on optical-field-induced current in a wide-gap dielectric. *New J. Phys.* **15,** 063019 (2013).
22. Vampa, G. *et al.* Theoretical analysis of high-harmonic generation in solids. *Phys. Rev. Lett.* **113,** 073901 (2014).
23. Hawkins, P. G., Ivanov, M. Y. & Yakovlev, V. S. Effect of multiple conduction bands on high-harmonic emission from dielectrics. *Phys. Rev. A* **91,** 013405 (2015).
24. Chin, A. H., Calderón, O. G. & Kono, J. Extreme midinfrared nonlinear optics in semiconductors. *Phys. Rev. Lett.* **86,** 3292–3295 (2001).
25. Ghimire, S. *et al.* Generation and propagation of high-order harmonics in crystals. *Phys. Rev. A* **85,** 043836 (2012).
26. Sekikawa, T., Katsura, T., Miura, S. & Watanabe, S. Measurement of the intensity-dependent atomic dipole phase of a high harmonic by frequency resolved optical gating. *Phys. Rev. Lett.* **88,** 193902 (2002).
27. Kemper, A. F., Moritz, B., Freericks, J. K. & Devereaux, T. P. Theoretical description of high-order harmonic generation in solids. *New J. Phys.* **15,** 023003 (2013).
28. Zener, C. A. Theory of the electrical breakdown of solid dielectrics. *Proc. R. Soc. Lond. A* **145,** 523–529 (1934).
29. Zhao, H., Loren, E. J., van Driel, H. M. & Smirl, A. L. Coherence control of Hall charge and spin currents. *Phys. Rev. Lett.* **96,** 246601 (2006).
30. Ladd, T. D. *et al.* Quantum computers. *Nature* **464,** 45–53 (2010).

Acknowledgements The work in Regensburg was supported by the European Research Council through grant no. 305003 (QUANTUMsubCYCLE) as well as by the Deutsche Forschungsgemeinschaft (through grant number HU 1598/2-1) and the work in Marburg by the Deutsche Forschungsgemeinschaft (through SFB 1083 and grant number KI 917/2-2).

Author Contributions M.H., F.L., O.S., U.H., S.W.K., M. Kira and R.H. conceived the study. M.H., F.L., O.S., M. Knorr and R.H. carried out the experiment and analysed the data. U.H., S.W.K. and M. Kira developed the quantum-mechanical model and carried out the computations. M.H., F.L., O.S., U.H., S.W.K., M. Kira and R.H. wrote the manuscript. All authors discussed the results.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.H. (rupert.huber@physik.uni-regensburg.de) or M. Kira (mackillo.kira@physik.uni-marburg.de).

METHODS

Experimental setup. A femtosecond titanium–sapphire laser amplifier (repetition rate, 3 kHz; pulse energy, 5.5 mJ; pulse duration, 33 fs; centre wavelength, 805 nm) is used to pump two parallel dual-stage optical parametric amplifiers tunable between centre wavelengths of 1.1 μm and 1.8 μm which deliver signal pulse energies of up to 0.5 mJ each. We generate phase-locked multi-terahertz pulses via difference frequency generation between these spectrally detuned near-infrared pulse trains[31].

High-order harmonics (HHs) are obtained by focusing these intense waveforms onto a gallium selenide (GaSe) crystal. Using samples 220 μm thick, power conversion efficiencies of approximately 7% for the whole HH spectrum covering the spectral range from 45 to 675 THz are observed. Resulting pulse energies of roughly 350 nJ are measured for a peak driving field of 72 MV cm^{-1} close to the observed damage threshold of GaSe. The threshold decreases for thinner samples suggesting thermal heating as the relevant damage mechanism. For the time-resolved experiments, a 60-μm-thick free-standing GaSe crystal is irradiated (centre frequency, 33 THz; electric peak field, 47 MV cm^{-1}; see inset Extended Data Fig. 1) under normal incidence to avoid phase matching effects. Under these conditions, the Bloch period amounts to 5.2 fs, corresponding to an energy of 0.8 eV. The polarization of the driving field points along the Γ–K direction of the hexagonal Brillouin zone of GaSe. The resulting spectrum (Extended Data Fig. 1) is recorded by means of a monochromator in combination with a calibrated pyroelectric detector (PED) and a lead sulfide (PbS) diode, as well as spectrometers with indium gallium arsenide (InGaAs) and cooled silicon (Si) detector arrays, respectively.

A YAG-based supercontinuum source with a combination of chirped mirrors and a prism compressor provides 8-fs gating pulses with a centre wavelength of 840 nm (Extended Data Fig. 3). Both the gating and the HH pulses are focused onto a 10-μm-thick β-BaB$_2$O$_4$ (BBO) crystal, which mediates two distinct nonlinear optical processes simultaneously, as follows. (1) The HHs are mixed with the gate by ultrabroadband sum-frequency generation. This cross-correlation signal is recorded as a function of the delay time t between the HH and gate pulses and encodes the temporal structure of both. (2) The terahertz fundamental wave induces an electro-optic polarization rotation of the gate pulse. This signature $S_{\mathrm{EOS\text{-}BBO}}(t)$ serves as a time marker directly linking the emitted HH intensity $I_{\mathrm{HH}}(t)$ with the generating terahertz field $E_{\mathrm{THz}}(t)$ (see below for details).

Determination of the absolute timescale. For a complete time-domain picture of high-harmonic generation (HHG) in solids, the emitted signal (E_{HH}) has to be measured and temporally correlated with the driving field (E_{THz}) with a precision significantly better than one optical cycle of E_{THz}. The most accurate way to determine the relative timing is to detect the co-propagating terahertz wave and the generated HHs simultaneously within the same detector. Additionally, the detector response as well as all propagation effects between HH generation and detection of the waveforms has to be taken into account.

In our experiment, we superimpose the HH waveform including the fundamental terahertz field and an ultrashort near-infrared pulse within the same 10-μm-thick BBO crystal to generate sum-frequency (S_{SF}) and electro-optic signals ($S_{\mathrm{EOS\text{-}BBO}}$). This electro-optic trace is employed as a temporal reference marker only. The strong material dispersion of BBO[32] in the multi-terahertz spectral region distorts the fundamental waveform. For a faithful determination of the temporal profile of the terahertz field, we exchange the BBO crystal with a 6.5-μm-thick ⟨110⟩-oriented zinc telluride (ZnTe) sampling crystal glued on a 300-μm-thick electro-optically inactive ZnTe substrate and record the electro-optic signal ($S_{\mathrm{EOS\text{-}ZnTe}}$) for different delay times (Extended Data Fig. 2). Via an analysis of the Gouy phase shift[33] for several recordings ('z-scan'), we ensure that the ZnTe crystal is placed in exactly the same position as the BBO crystal.

The accuracy of this correlation procedure is estimated to be 1.2 fs by repeatedly exchanging the BBO and ZnTe crystals and calculating the standard deviation of the temporal delay between the transients recorded in BBO and ZnTe detectors, respectively. The error resulting from a misalignment of the detection crystals is on the order of 0.1 fs and may be neglected.

Finally, we correct for effects of phase-matching and dispersion of the $\chi^{(2)}$ nonlinearity using standard procedures[34] (compare Extended Data Fig. 2b) to obtain the terahertz field E_{Det} at the detector position for a given signal $S_{\mathrm{EOS\text{-}ZnTe}}$. The thickness of the ZnTe crystal d_{ZnTe} is confirmed by optical interference to be $d_{\mathrm{ZnTe}} = 6.5\ \mu\mathrm{m} \pm 0.7\ \mu\mathrm{m}$, resulting in an error of the peak field positions of 0.9 fs due to the uncertainty in the determination of the crystal thickness. Consequently, the relative error in the temporal delay between the sum-frequency signal (S_{SF}) generated by the emitted HHs and the terahertz field at the detection focus is 1.5 fs, that is, approximately 20 times shorter than one oscillation period of the terahertz waveform.

Propagation effects between HHG and detection of the fields may modify the measured signals additionally. We employ only reflective optics between the generation and detection stages. Consequently, differences in group velocities are negligible in our experiment. The GaSe sample is placed directly in the focus of the terahertz wave, leading to an additional Gouy phase[33] offset of π/2. For our focusing conditions, the Rayleigh length of the driving terahertz beam is about 500 μm, which is much longer than the crystal thickness of 60 μm and allows us to neglect the uncertainty in the sample position. The Gouy phase shift at the detection focus is also fully accounted for.

During propagation inside the sample, dispersion of the HH group velocity would lead to a relative delay of the spectral components corresponding to different harmonic orders. Since we deliberately avoid phase-matching during HHG in GaSe, the emission of HHs is confined to a thin region near the back facet of the GaSe crystal with a coherence length of, for example, 6 μm for the 9th harmonic, and no substantial chirp due to linear material dispersion is observed in the experiment. Note that the rear facet itself does not have an essential role in HHG. The HH radiation rather originates from the bulk of the crystal, as seen from the fact that phase-matching effects occur in thicker samples[5].

Double-blind XFROG algorithm. For the characterization of the temporal structure of terahertz HHs, we perform cross-correlation frequency resolved optical gating[35] (XFROG) between an 8-fs gating pulse and the HH pulse train. Our analysis is based on standard XFROG algorithms[36,37], customized to take into account the ultrabroadband nature of the HH spectra. The gating pulse encompasses a spectral bandwidth of more than 100 THz (Extended Data Fig. 3), while the emitted HH spectrum continuously covers multiple optical octaves. As a consequence, ultrabroadband sum-frequency mixing signals comprise a frequency range of more than 300 THz (Extended Data Fig. 3), benefitting from the huge acceptance bandwidth of the 10-μm-thick BBO crystal. Despite the enormous bandwidths, the XFROG algorithm reliably converges for discretionary runs and data sets and reconstructs the measured spectrograms to a very high degree of congruency (compare Fig. 1). The robustness of the scheme is additionally confirmed by comparison of the retrieved temporal shape of the gating pulse to an independent SHG FROG measurement, reconstructed using a separate algorithm[38] (Extended Data Fig. 3b).

The robust XFROG reconstruction allows us to retrieve the subcycle profile of both the intensity envelope $I_{\mathrm{HH}}(t)$ and the relative phase $\phi_{\mathrm{HH}}(t)$ of the HH bursts. While the temporal structure of $I_{\mathrm{HH}}(t)$ is shown in Fig. 1e, Extended Data Fig. 7 depicts the relative phase. The modulations of the phase are relatively small as already indicated by the almost simultaneous appearance of all sum-frequency spectral contributions (Fig. 1c). To study the phase evolution in greater detail, we derive the instantaneous frequency $\nu_{\mathrm{i}} = (2\pi)^{-1}\partial\phi_{\mathrm{HH}}/\partial t$ by numerical differentiation. The instantaneous frequency is ramped up during the rising edge of a HH burst, peaks together with the intensity envelope and decreases again (Extended Data Fig. 7). This slight double-chirp is an indicator of Bloch-type acceleration of carriers within the conduction band. Furthermore, the phase retrieval allows for the reconstruction of the HH field trace (compare red waveform in Fig. 1a).

Quantum many-body model. We use the HHG theory developed in refs 20 and 39 to describe the coherent interplay of interband excitations and intraband currents and apply this microscopic model to study the time resolved emission of GaSe excited with extremely strong terahertz fields. The numerical calculations are based on a one-dimensional five-band model as depicted in Extended Data Fig. 4a, including two conduction bands ($\lambda = \mathrm{e}_1, \mathrm{e}_2$, blue lines) and three valence bands ($\lambda = \mathrm{h}_1, \mathrm{h}_2, \mathrm{h}_3$, red lines). The individual bands are modelled by effectively one-dimensional tight-binding bands[39], with material parameters taken from ref. 40. In the following, we will review relevant details, while more comprehensive derivations can be found in ref. 5.

The presence of a sufficiently strong external electric field $E(t)$ induces interband transitions where electrons are transferred between two different bands λ and λ', as schematically depicted by the blue spheres and the black arrow in Extended Data Fig. 4b. The dipole-matrix element $d_{\lambda\lambda'}$ determines the strength of the transition. The inter-valence-band dipole-matrix element $d_{\mathrm{h}_1\mathrm{h}_2}$ exceeds the values of $d_{\mathrm{h}_1\mathrm{e}_1}$ and $d_{\mathrm{h}_2\mathrm{e}_1}$ by one order of magnitude[41]. The strongest driving fields applied in the calculations correspond to peak Rabi energies of 1.5 eV for inter-valence-band transitions and 0.12 eV for valence-to-conduction-band transitions. In addition to the interband transitions, $E(t)$ also drives intraband currents where electrons and holes are accelerated in their respective bands[4–8,20–23,25,27,42], see Extended Data Fig. 4c.

Defining electron and hole occupations f_k^{λ} and microscopic polarizations $p_k^{\lambda\lambda'}$ for $\lambda \neq \lambda'$, the well-known semiconductor Bloch equations[43] (SBEs) describe the time evolution of polarizations and carrier occupations. The SBEs read

$$\begin{aligned}\hbar\frac{\partial}{\partial t}f_k^{\mathrm{e}_1} = -2\mathrm{Im}\left[d_{\mathrm{e}_1\mathrm{e}_2}(k)E(t)p_k^{\mathrm{e}_2\mathrm{e}_1} + \sum_{\mathrm{h}_\lambda} d_{\mathrm{e}_1\mathrm{h}_\lambda}(k)E(t)\left(p_k^{\mathrm{h}_\lambda\mathrm{e}_1}\right)^*\right] \\ + |e|E(t)\nabla_k f_k^{\mathrm{e}_1} + \hbar\frac{\partial}{\partial t}f_k^{\mathrm{e}_1}\Big|_{\mathrm{relax}}\end{aligned} \qquad (1)$$

for the carrier occupation of the first conduction band, and

$$i\hbar\frac{\partial}{\partial t}p_k^{\mathrm{h}_i\mathrm{e}_j}=\left(\varepsilon_k^{\mathrm{e}_j}+\varepsilon_k^{\mathrm{h}_i}-i\frac{\hbar}{T_2}\right)p_k^{\mathrm{h}_i\mathrm{e}_j}-d_{\mathrm{e}_j\mathrm{h}_i}(k)E(t)\left(1-f_k^{\mathrm{e}_j}-f_k^{\mathrm{h}_i}\right)+i|e|E(t)\nabla_k p_k^{\mathrm{h}_i\mathrm{e}_j}$$
$$+E(t)\sum_{\mathrm{e}_\lambda\neq\mathrm{e}_j}\left[d_{\mathrm{e}_\lambda\mathrm{h}_i}(k)p_k^{\mathrm{e}_\lambda\mathrm{e}_j}-d_{\mathrm{e}_j\mathrm{e}_\lambda}(k)p_k^{\mathrm{h}_i\mathrm{e}_\lambda}\right] \quad (2)$$
$$+E(t)\sum_{\mathrm{h}_\lambda\neq\mathrm{h}_i}\left[d_{\mathrm{h}_\lambda\mathrm{h}_i}(k)p_k^{\mathrm{h}_\lambda\mathrm{e}_j}-d_{\mathrm{e}_j\mathrm{h}_\lambda}(k)p_k^{\mathrm{h}_i\mathrm{h}_\lambda}\right]$$

for the microscopic polarizations between a valence band and a conduction band. In order to capture the dominant effects of the higher-order correlations, we include a phenomenological dephasing via the decay time $T_2 = 1.1$ fs in the polarization dynamics, following refs 5 and 44.

Multi-photon coherence is particularly susceptible to dephasing because any scattering scrambles its phase relations. In fact, similarly fast decay times of coherences have been reported in different systems ranging from bandgap dynamics in silicon[45] (0.5-fs decay) via general HHG in solids (ref. 22 (4-fs decay) and ref. 46 (sub-10-fs decay)) to atomic HHG[47] (sub-10-fs decay). This ultrafast timescale results from electron–electron scattering[45] as well as polarization–polarization scattering[47], which is fostered by the broad carrier and polarization distribution created in HHG. In addition, the terahertz field excites a large amount of electrons much faster than Coulomb screening builds up. The unscreened Coulomb interaction enhances scattering rates significantly, as discussed for example, in ref. 47. The resulting dynamics are numerically solved for an initially unexcited system using an excitation field $E(t)$ modelled closely to the experimental terahertz waveform.

The emission intensity of a coherently excited semiconductor consists of a polarization source $P(t)$ and a current source $J(t)$,

$$P(t)=\sum_{\lambda,\lambda',k}d_{\lambda,\lambda'}(k)p_k^{\lambda,\lambda'} \text{ and } J(t)=\sum_{\lambda,k}j_\lambda(k)f_k^\lambda \quad (3)$$

with the current matrix element $j_\lambda(k)=\frac{|e|}{\hbar}\nabla_k\varepsilon_k^\lambda$.

To account for the damping of currents in a realistic system, a phenomenological carrier relaxation $\hbar\frac{\partial}{\partial t}f_k^\lambda\big|_{\mathrm{relax.}}=-\frac{1}{\tau}f_{k,A}^\lambda$ is included, which damps the antisymmetric part $f_{k,A}^\lambda=\frac{1}{2}\left[f_k^\lambda-f_{(-k)}^\lambda\right]$ of the carrier distributions, producing the correct decay for the currents. We apply this relaxation to all valence bands and the first conduction band, with the relaxation time $\tau = 7$ fs. Owing to its flat shape, the second conduction band does not contribute essentially to the currents.

The emission intensity is defined by the total effective current given by the sum of intraband currents $J(t)$ and the rate of change of the macroscopic polarization[48] $\frac{\partial}{\partial t}P(t)$,

$$E_{\mathrm{HHG}}(t)\propto\frac{\partial}{\partial t}P(t)+J(t) \quad (4)$$

The envelope and phase of the numerically computed time trace is extracted via a Hilbert transform[49]. As experimental evidence from phase matching effects (see 'Determination of the absolute timescale' in Methods) suggests, surface effects do not contribute considerably to the HHG and are, thus, neglected in our model.

Strong-field quantum interference. As described in the previous section, all our numerical analysis is performed with the five-band model. However, in order to obtain an intuitive understanding of the dominant quantum-interference paths contributing to the experimental observations, it is sufficient to consider a simplified system that includes only the three most relevant bands. As depicted in Extended Data Fig. 5a, we use two valence bands h_1 and h_2 (red solid lines) and one conduction band e (blue solid line). The transition probabilities between the bands are given by the magnitude of the dipole-matrix elements d_{eh_1}, d_{eh_2} and $d_{\mathrm{h}_2\mathrm{h}_1}$ (black arrows). For an initially unexcited system, only transitions from the two valence bands h_1 and h_2 to the conduction band e are possible, as both valence bands are completely filled, preventing any transition between them.

In order to study the different excitation paths leading to the effective transition $\mathrm{h}_1 \rightarrow \mathrm{e}$, the sources

$$i\hbar\frac{\partial}{\partial t}p_k^{\mathrm{h}_1\mathrm{e}}\bigg|_{\mathrm{ex.}}=-d_{\mathrm{eh}_1}(k)E(t)(1-f_k^e-f_k^{\mathrm{h}_1})+d_{\mathrm{h}_2\mathrm{h}_1}(k)E(t)p_k^{\mathrm{h}_2\mathrm{e}}+d_{\mathrm{eh}_2}(k)E(t)p_k^{\mathrm{h}_1\mathrm{h}_2} \quad (5)$$

based on equation (2) are analysed for the situation of three bands depicted in Extended Data Fig. 5a. The gradient term is omitted here, because it only redistributes the carrier momentum inside a band without inducing transitions between different bands. For the sake of a simpler notation, we will also omit the explicit k-dependence from now on.

The first excitation path $\mathrm{h}_1 \rightarrow \mathrm{e}$ (first term in equation (5)) is depicted by the blue arrow in Extended Data Fig. 5b. The terahertz driving field creates a polarization between the first valence and the conduction band, which is proportional to the driving field strength and the corresponding dipole-matrix element. We call this the direct excitation path, $p_k^{\mathrm{h}_1\mathrm{e}}|_{\mathrm{direct}}\propto d_{\mathrm{eh}_1}E(t)$; it is initially linear in the electric field $E(t)$ and thus of odd order with respect to the driving field sign. For strong fields, this polarization will eventually excite electrons by multi-photon absorption to the conduction band. This change in the carrier occupations given by $f_k^{\mathrm{e}}|_{\mathrm{direct}}\propto d_{\mathrm{eh}_1}E(t)(p_k^{\mathrm{h}_1\mathrm{e}}|_{\mathrm{direct}})^*\propto d_{\mathrm{eh}_1}^2E(t)^2$ will modulate the initially linear dependence via the electron occupations of the involved states. Nevertheless the odd symmetry of the transition is preserved: Due to the mutual coupling of polarization and carrier occupations, that is, $p_k^{\mathrm{h}_1\mathrm{e}}|_{\mathrm{direct}}\propto d_{\mathrm{eh}_1}E(t)-d_{\mathrm{eh}_1}E(t)f_k^{\mathrm{e}}|_{\mathrm{direct}}$ and $f_k^{\mathrm{e}}|_{\mathrm{direct}}\propto d_{\mathrm{eh}_1}E(t)(p_k^{\mathrm{h}_1\mathrm{e}}|_{\mathrm{direct}})^*\propto d_{\mathrm{eh}_1}^2E(t)^2$ the initially linear dependence will be replaced by a series of terms, which are all of odd order in the driving field, that is, $p_k^{\mathrm{h}_1\mathrm{e}}|_{\mathrm{direct}}\propto d_{\mathrm{eh}_1}E(t)-d_{\mathrm{eh}_1}^3E(t)^3+\cdots$.

In the same way, the excitation path $\mathrm{h}_2 \rightarrow \mathrm{e}$ will also be driven by the electric field (three-dimensional red arrow in Extended Data Fig. 5c). Hence, $p_k^{\mathrm{h}_2\mathrm{e}}\propto d_{\mathrm{eh}_2}E(t)$ initially has a linear dependence on the driving field. Eventually electrons from h_2 will also be excited to the conduction band, which will create vacancies in h_2 and allow for transitions $\mathrm{h}_1 \rightarrow \mathrm{h}_2$ between the valence bands.

As soon as $\mathrm{h}_1 \rightarrow \mathrm{h}_2$ transitions (flat red arrow in Extended Data Fig. 5c) become possible, the second excitation path, described by the second term in equation (5), $p_k^{\mathrm{h}_1\mathrm{e}}|_{\mathrm{second}}\propto d_{\mathrm{h}_2\mathrm{h}_1}E(t)p_k^{\mathrm{h}_2\mathrm{e}}$, opens up. This path leads to a field induced excitation from the first valence band via the second valence band to the conduction band and is mediated by the polarization $p_k^{\mathrm{h}_2\mathrm{e}}$. Thus only if a polarization $p_k^{\mathrm{h}_2\mathrm{e}}$ is already present and the transition $\mathrm{h}_1 \rightarrow \mathrm{h}_2$ is possible, this indirect path exists. However, $p_k^{\mathrm{h}_2\mathrm{e}}$ itself depends initially linearly on the electric field. Therefore, the indirect path $p_k^{\mathrm{h}_1\mathrm{e}}|_{\mathrm{indirect}}\propto d_{\mathrm{h}_2\mathrm{h}_1}d_{\mathrm{eh}_2}E(t)^2$ depends quadratically on the electric field and is proportional to the product of the transition dipole-matrix elements $d_{\mathrm{h}_2\mathrm{h}_1}$ for $\mathrm{h}_1 \rightarrow \mathrm{h}_2$ and d_{eh_2} for $\mathrm{h}_2 \rightarrow \mathrm{e}$. For strong excitations this quadratic dependence is replaced by a series of terms which are of even order in the driving field. Overall, the transition $\mathrm{h}_1 \rightarrow \mathrm{h}_2 \rightarrow \mathrm{e}$ has an even symmetry with respect to the driving field.

Combining direct and indirect paths, the sum of their amplitudes—not their probabilities—defines the total transition yield. Therefore, in the lowest order, the total polarization

$$p_k^{\mathrm{h}_1\mathrm{e}}|_{\mathrm{total}}=p_k^{\mathrm{h}_1\mathrm{e}}|_{\mathrm{direct}}+p_k^{\mathrm{h}_1\mathrm{e}}|_{\mathrm{indirect}}\propto d_{\mathrm{eh}_1}E(t)+d_{\mathrm{h}_2\mathrm{h}_1}d_{\mathrm{eh}_2}E(t)^2 \quad (6)$$

consists of linear and quadratic contributions. For strong excitations, the field dependence is replaced by functions which are of even and odd order in the driving field. The resulting quantum interference controls the symmetry of the excitation $\mathrm{h}_1 \rightarrow \mathrm{e}$ because direct and indirect paths involve different field orders, as depicted by the red and blue arrows in Extended Data Fig. 5d. Therefore, changing the sign or the phase of the driving field $E(t)$ allows for direct control of the quantum properties of the excited state e. Equation (6) shows that all three transitions $\mathrm{h}_1 \rightarrow \mathrm{e}$, $\mathrm{h}_2 \rightarrow \mathrm{e}$, and $\mathrm{h}_1 \rightarrow \mathrm{h}_2$ are needed to produce quantum interference. If one transition $\lambda \rightarrow \lambda'$ is forbidden, that is, $d_{\lambda\lambda'}=0$, $p_k^{\mathrm{h}_1\mathrm{e}}|_{\mathrm{total}}$ can only have one contribution with either even or odd parity in E.

Another contribution to the indirect path $\mathrm{h}_1 \rightarrow \mathrm{h}_2 \rightarrow \mathrm{e}$ is described by the third term in equation (5), $p_k^{\mathrm{h}_1\mathrm{e}}|_{\mathrm{third}}=d_{\mathrm{eh}_2}(k)E(t)p_k^{\mathrm{h}_1\mathrm{h}_2}$. Here, the polarization between the two valence bands mediates the transition to the conduction band. However, $p_k^{\mathrm{h}_1\mathrm{h}_2}$ is not driven by the electric field directly and may thus have a different field dependence compared to the $p_k^{\mathrm{h}_2\mathrm{e}}$-mediated transitions, effectively creating an additional excitation path.

In the full five-band system, the additional bands will greatly increase the number of possible excitation paths. Even transitions involving three bands, for example, $\mathrm{h}_1 \rightarrow \mathrm{h}_2 \rightarrow \mathrm{h}_3 \rightarrow \mathrm{e}$, can become possible. The resulting excitation path will also contribute to the transition $\mathrm{h}_1 \rightarrow \mathrm{e}$ analysed above, leading to a quantum interference of at least three different excitation paths, with a possibly even more complicated field dependence. In the same way, excitations where the second valence band is the initial or final state, that is, $\mathrm{h}_2 \rightarrow \mathrm{e}$ and $\mathrm{h}_2 \rightarrow \mathrm{h}_1$, also contribute to the emission (direct paths).

Perturbative versus non-perturbative quantum interference. In principle, quantum interference could also be created in the perturbative regime. However, the frequency and intensity of the driving field would have to be chosen carefully to balance the involved excitation paths that depend on different orders of the driving field. A slight change in the properties of the pump field would then lead to a preferred path and, as a consequence, destroy the quantum interference effect since that preferred path would mainly contribute to the emission. As shown in the main text, the unipolar emission as a sign of quantum interference is observed for a variety of driving field strengths and central frequencies, which is a strong indicator for non-perturbative coherent control.

Owing to the very strong terahertz fields used in the experiments, the transition amplitudes are modulated by the electron occupations f_k^λ of the involved states. The resulting massive dynamic occupation changes are beyond the validity of the perturbative analysis and the non-perturbative regime is reached. Technically, the power-law dependence of transition amplitudes on the electric field E is replaced by some nonlinear functions, denoted by $A_o(E)$ and $A_e(E)$ for direct excitations (for example, $h_1 \rightarrow e$) and indirect paths (for example, $h_1 \rightarrow h_2 \rightarrow e$), respectively. Nevertheless, as in the perturbative analysis, we can still identify an odd symmetry $A_o(-E) = -A_o(E)$ for the direct path and an even symmetry $A_e(-E) = A_e(E)$ for the indirect path. Also the superposition principle remains valid, implying proportionality $A_e(E) + A_o(E)$ for the total transition amplitude composed of the two excitation paths. The sign of the field thus controls the total outcome of the excitations, yielding $A_e(|E|) + A_o(|E|)$ for positive E and $A_e(|E|) - A_o(|E|)$ for negative E. Therefore, the system described by equation (5) inherently contains non-perturbative coherent control via quantum interference. Note that, while being assigned direct and indirect paths, the excitation paths are true multi-photon transitions in the non-perturbative regime, not perturbative single and two-photon transitions.

Interference path efficiency. To determine the precise weight of direct versus indirect path contributions, we compare the total electron density, n_{e_1}, generated in the first conduction band with ($F_{cc} = 1$) and without ($F_{cc} = 0$) the indirect paths. We define n_{e_1} 300 fs after the terahertz field, that is, at a time when most of the carrier generation has been completed. We then construct the interference path efficiency

$$\eta_{\mathrm{IPE}} \equiv \frac{n_{e_1}(F_{cc}=1) - n_{e_1}(F_{cc}=0)}{n_{e_1}(F_{cc}=0)} \tag{7}$$

by computing the ratio of excess electrons created by the indirect paths ($n_{e_1}(F_{cc}=1) - n_{e_1}(F_{cc}=0)$) and only by the direct paths ($n_{e_1}(F_{cc}=0)$).

Extended Data Fig. 5e shows η_{IPE} as a function of the terahertz field strength. For low E_{THz}, η_{IPE} remains close to zero, indicating the dominance of the direct paths. Increasing the external driving field strength to 22 MV cm^{-1} elevates η_{IPE} to 50%. For $E_{\mathrm{THz}} = 30$ MV cm^{-1}, the direct and the indirect paths reach the same efficiency ($\eta_{\mathrm{IPE}} = 1$). In other words, non-perturbative excitations tend to balance the relative weights of the excitation paths, making interference effects strong, as described in the main text. We even observe that η_{IPE} starts to decrease slightly, indicating a Rabi-flopping-type saturation, which is a unique hallmark of strongly non-perturbative excitations.

Coherent control level. Extended Data Fig. 6 shows the computed HH intensity envelopes as a function of the coherent control factor F_{cc} as presented in Fig. 2c. However, all time traces are now displayed on the same absolute intensity scale. Including all transition paths ($F_{cc} = 1$) enhances the emission intensity by roughly 30 times, that is, more than simple two-path interference (maximum enhancement by a factor of four) predicts. This is indeed expected since switching off indirect transition channels ($F_{cc} = 0$) eliminates the contribution of multiple polarization combinations in the emission intensity, which lowers the HHG efficiency drastically. Additionally, GaSe has a particularly strong dipole matrix-element between the hole bands, as explained in ref. 41. This contribution is also missing when F_{cc} is set to zero, together with multiple interference pathways. By closing the indirect transition channels in the $F_{cc} = 0$ calculation, the number of excited carriers is also reduced by a factor of two compared to the full calculation ($F_{cc} = 1$). As a combination of all these effects, $F_{cc} = 0$ yields a significant reduction in HHG, besides the qualitative difference in temporal emission, compared to the full computation.

Minimal requirements for non-perturbative quantum interference. In the main text, we have identified the following list of minimal conditions for the non-perturbative quantum interference:

(1) The system must have at least three states that are all mutually dipole coupled. The dipole transitions can then be illustrated in a triangle diagram, as shown in Extended Data Fig. 10a.

(2) The terahertz field must be non-resonant with each of the transitions because the excitation channels can otherwise not be balanced.

(3) The terahertz field must be strong enough to generate non-perturbative excitations where populations change within less than one oscillation cycle of the exciting field.

In the schematic of Extended Data Fig. 10a, circles denote three states (e, 1, and 2) coupled with three dipole-allowed paths (arrows). The actual dipole moment between any two states λ and λ' follows from

$$\boldsymbol{d}_{\lambda\lambda'} \equiv \frac{1}{\Omega}\int_{\Omega} \mathrm{d}r^3 \phi_\lambda^*(\boldsymbol{r}) e\boldsymbol{r} \phi_{\lambda'}(\boldsymbol{r}) \tag{8}$$

where $\phi_\lambda(\boldsymbol{r})$ is the single-particle wave function of a given state λ. For solids, the integral can be performed over the unit-cell volume Ω.

An inversion symmetric potential, depicted schematically in Extended Data Fig. 10b, produces single-particle eigenfunctions that have either even ($\phi_\lambda(-\boldsymbol{r}) = \phi_\lambda(\boldsymbol{r})$) or odd ($\phi_\lambda(-\boldsymbol{r}) = -\phi_\lambda(\boldsymbol{r})$) parity. Inserting these into equation (8), $\boldsymbol{d}_{\lambda\lambda'}$ vanishes when both λ and λ' have the same parity and exists only between states with different parity. Since at least two out of the three states must have the same parity, an inversion symmetric system cannot form the required triangle system, making only Λ-, V- or ladder-transitions possible. Hence, the discovered non-perturbative quantum interference cannot be observed in inversion symmetric molecules or solids. More generally, inversion symmetric systems can have closed transition loops only among an even number of states, which always creates a total transition with odd symmetry with respect to the driving field. Inversion symmetric systems can therefore not exhibit the observed non-perturbative quantum interference even via more complex excitation paths. When the potential does not possess inversion symmetry as illustrated in Extended Data Fig. 10c, $\phi_\lambda(\boldsymbol{r})$ has neither even nor odd symmetry. Owing to the indefinite parity, $\boldsymbol{d}_{\lambda\lambda'}$ can exist between any combination of states, as is the case in GaSe. Solids or molecules lacking inversion symmetry can, thus, exhibit three states which are dipole-connected in a triangle configuration.

In Extended Data Fig. 10a, non-perturbative quantum interference can be realized in many different ways. One possibility is to choose $1 \rightarrow e$ to be the direct transition and $1 \rightarrow 2 \rightarrow e$ the indirect path (scenario QI1). Equivalently, the direct transition $2 \rightarrow e$ can interfere with the indirect path $2 \rightarrow 1 \rightarrow e$ (scenario QI2). To determine the interference conditions in these two cases, we apply $\phi_\lambda \rightarrow \tilde{\phi}_\lambda e^{i\theta_\lambda}$ in equation (8) and get $\tilde{d}_{\lambda\lambda'}$ with states $\tilde{\phi}_\lambda$ and $d_{\lambda\lambda'} = e^{i(\theta_{\lambda'} - \theta_\lambda)}\tilde{d}_{\lambda\lambda'}$. We choose θ_λ to produce real valued $\tilde{d}_{e1} = |\tilde{d}_{e1}|$ and $\tilde{d}_{e2} = |\tilde{d}_{e2}|$ while $\tilde{d}_{21}$ remains complex valued. As derived in equation (6), the overall strength of the QI1 transition scales with

$$|p_k^{1e}|_{\mathrm{QI1}} \propto |\tilde{d}_{e1}E(t) + \tilde{d}_{21}\tilde{d}_{e2}E(t)^2| = \tilde{d}_{e1}|E(t) + zE(t)^2| \tag{9}$$

with $z \equiv \left|\frac{\tilde{d}_{e2}}{\tilde{d}_{e1}}\right|\tilde{d}_{21}$. Similar analysis for QI2 produces the expression within the absolute value in equation (9) with $z = \left|\frac{\tilde{d}_{e1}}{\tilde{d}_{e2}}\right|\tilde{d}_{12}$. Hence, both scenarios QI1 and QI2 lead to constructive (destructive) interference for a positive (negative) sign of the driving field and thus qualitatively contribute in the same way to the overall outcome of HH emission.

Ultrashort pulse shaping. The new cross-correlation scheme allows us to determine the exact timing with respect to the driving terahertz waveform as well as the intensity of the emitted HH bursts. The relative weight of HH bursts emitted at different half-cycles of the phase-stable waveforms can be controlled via the carrier-envelope phase (CEP) of the transients[5,31]: while continuously varying the CEP of the driving field, we observe that the individual peaks of the HH pulse train shift under a fixed envelope (Extended Data Fig. 8), defined by the envelope of the driving waveform. Quantum interference between more than two energy bands causes a pronounced polarity dependence of HH emission, leading to a temporal spacing of the HH bursts by the full driving period T instead of $T/2$. This intrinsic property facilitates realization of isolated ultrashort field bursts even if the intensity envelope of the driving pulse is substantially longer than one half-cycle.

If the CEP changes by π (Extended Data Fig. 8b) one may switch between one dominant HH pulse centred at the maximum field position (black curve) and two equally intense bursts corresponding to two positive driving half-cycles of equal field strengths (red curve). The direct CEP-controllability may thus allow for the generation of isolated ultrashort HH bursts as proposed in theoretical studies[6–8].

The flat spectral phase within single emission bursts of HHs favours the emission of extremely short optical pulses directly from the bulk semiconductor. Based on a semiclassical intraband calculation (supplementary information of ref. 5), we estimate that well-chosen spectral amplitude filtering could allow for the generation of 3-fs short bursts (Extended Data Fig. 9) without external pulse compression. Extending the subcycle pulse shaping techniques introduced in this study to higher frequencies by using wide-gap semiconductors as HH emitters may thus ultimately yield even subfemtosecond bursts from solid-state sources.

31. Sell, A., Leitenstorfer, A. & Huber, R. Phase-locked generation and field-resolved detection of widely tunable terahertz pulses with amplitudes exceeding 100 MV/cm. *Opt. Lett.* **33,** 2767–2769 (2008).

32. Eimerl, D., Davis, L., Velsko, S., Graham, E. K. & Zalkin, A. Optical, mechanical, and thermal properties of barium borate. *J. Appl. Phys.* **62,** 1968–1983 (1987).

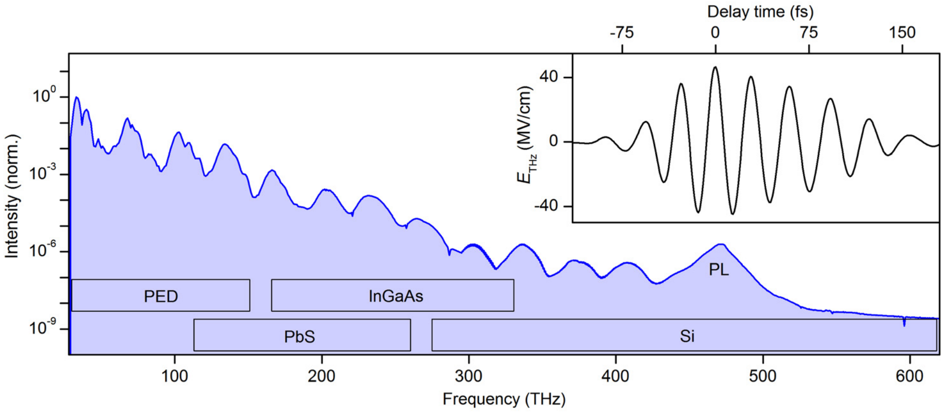
33. Ruffin, A. B., Rudd, J. V., Whitaker, J. F., Feng, S. & Winful, H. G. Direct observation of the Gouy phase shift with single-cycle terahertz pulses. *Phys. Rev. Lett.* **83,** 3410–3413 (1999).

34. Gallot, G. & Grischkowsky, D. Electro-optic detection of terahertz radiation. *J. Opt. Soc. Am. B* **16,** 1204–1212 (1999).

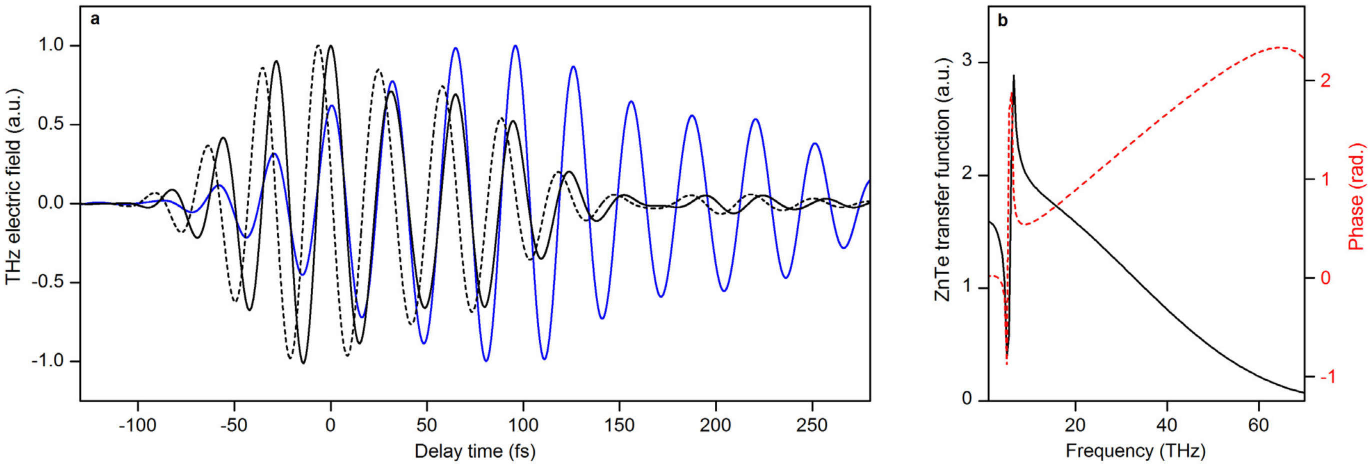
35. Linden, S., Giessen, H. & Kuhl, J. XFROG — A new method for amplitude and phase characterization of weak ultrashort pulses. *Phys. Status Solidi B* **206,** 119–124 (1998).

36. Kane, D. J. Real-time measurement of ultrashort laser pulses using principal component generalized projections. *IEEE J. Sel. Top. Quantum Electron.* **4,** 278–284 (1998).

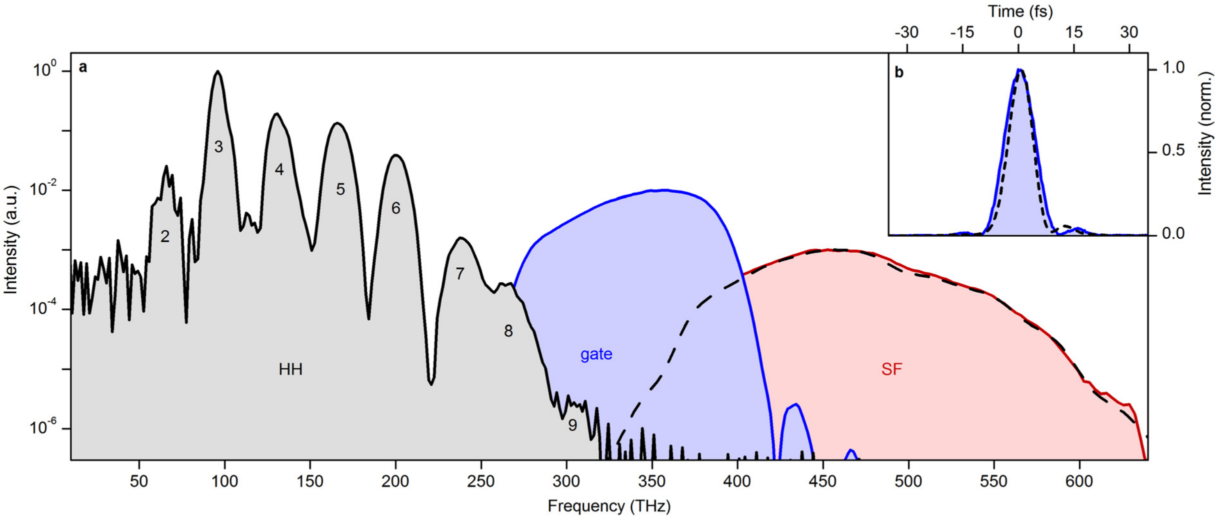
37. Wyatt, A. Frequency-resolved optical gating. http://www.mathworks.com/matlabcentral/fileexchange/16235-frequency-resolved-optical-gating-frog- (MATLAB central file exchange, 7 July 2008).
38. Kane, D. J. Recent progress toward real-time measurement of ultrashort laser pulses. *IEEE J. Quantum Electron.* **35,** 421–431 (1999).
39. Golde, D., Kira, M., Meier, T. & Koch, S. W. Microscopic theory of the extremely nonlinear terahertz response of semiconductors. *Phys. Status Solidi B* **248,** 863–866 (2011).
40. Schlüter, M. *et al.* Optical properties of GaSe and GaS_xSe_{1-x} mixed crystals. *Phys. Rev. B* **13,** 3534–3547 (1976).
41. Segura, A., Bouvier, J., Andrés, M. V., Manjón, F. J. & Muñoz, V. Strong optical nonlinearities in gallium and indium selenides related to inter-valence-band transitions induced by light pulses. *Phys. Rev. B* **56,** 4075–4084 (1997).
42. Moiseyev, N. Selection rules for harmonic generation in solids. *Phys. Rev. A* **91,** 053811 (2015).
43. Kira, M. & Koch, S. W. *Semiconductor Quantum Optics* (Cambridge Univ. Press, 2011).
44. Golde, D., Meier, T. & Koch, S. W. Microscopic analysis of extreme nonlinear optics in semiconductor nanostructures. *J. Opt. Soc. Am. B* **23,** 2559–2565 (2006).
45. Schultze, M. *et al.* Attosecond band-gap dynamics in silicon. *Science* **346,** 1348–1352 (2014).
46. Vu, Q. T. *et al.* Light-induced gaps in semiconductor band-to-band transitions. *Phys. Rev. Lett.* **92,** 217403 (2004).
47. Schuh, K., Hader, J., Moloney, J. V. & Koch, S. W. Influence of many-body interactions during the ionization of gases by short intense optical pulses. *Phys. Rev. E* **89,** 033103 (2014).
48. Kira, M., Jahnke, F., Hoyer, W. & Koch, S. W. Quantum theory of spontaneous emission and coherent effects in semiconductor microstructures. *Prog. Quantum Electron.* **23,** 189–279 (1999).
49. Liu, Y.-W. in *Fourier Transform Applications* (ed. Salih, S.) 291–300 (InTech, 2012).



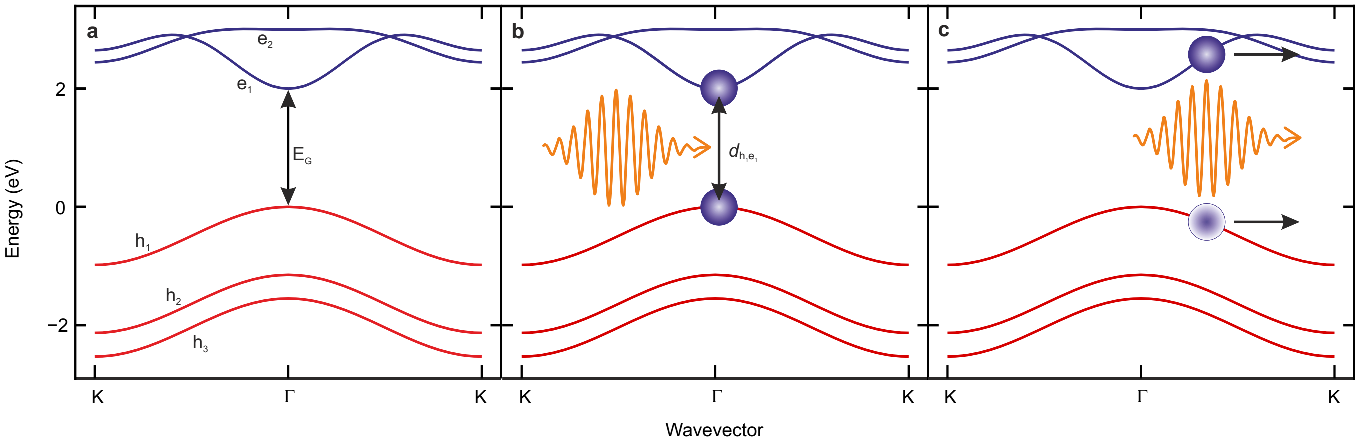
Extended Data Figure 1 | High-order harmonic (HH) spectrum generated by intense phase-locked terahertz pulses in bulk GaSe. The ultrabroadband HH emission is recorded using a monochromator with a calibrated pyroelectric detector (PED), a lead sulfide diode (PbS) and spectrometers employing InGaAs and cooled Si detectors. At a frequency of 476 THz interband photoluminescence (PL) dominates the spectrum. The driving multi-terahertz waveform is shown in the inset.



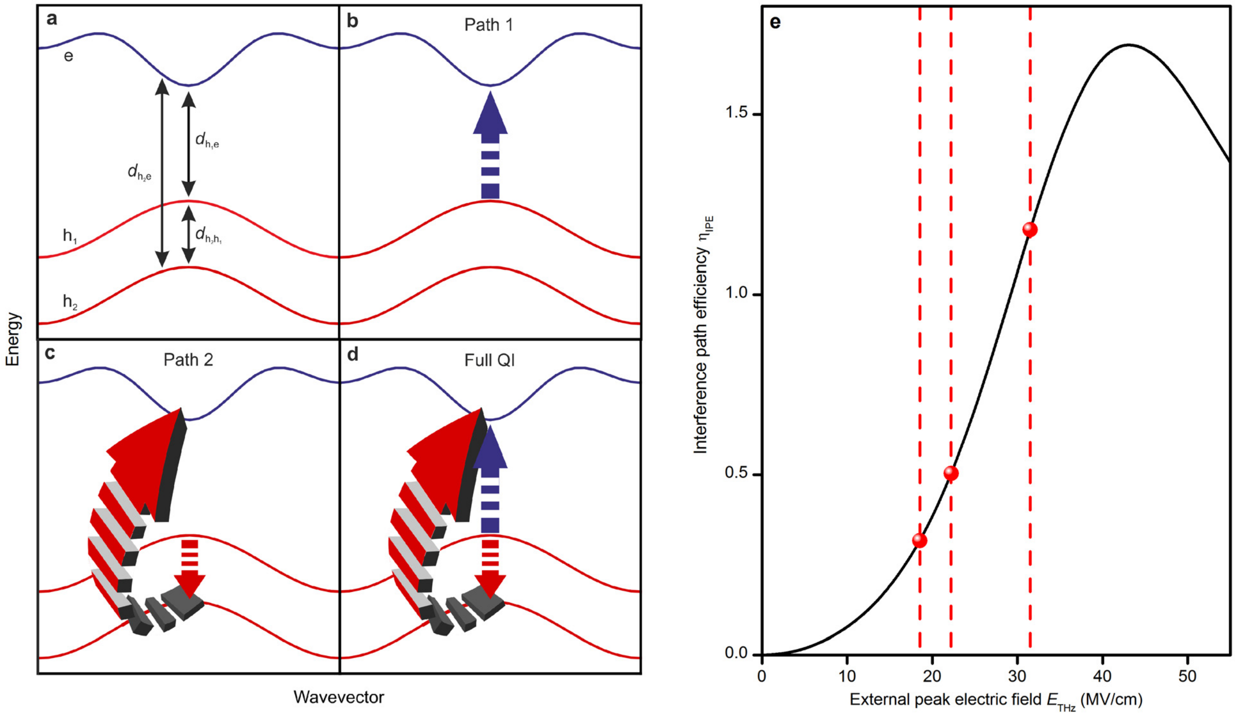
Extended Data Figure 2 | Determination of the absolute timescale. **a**, Electro-optic signal of the multi-terahertz driving field as recorded using a BBO detector (thickness, 10 μm; blue curve, $S_{\mathrm{EOS\text{-}BBO}}$) and a ⟨110⟩-oriented ZnTe crystal (thickness, 6.5 μm; black dashed curve, $S_{\mathrm{EOS\text{-}ZnTe}}$) for spectral components between 12 and 45 THz. The black solid curve represents the corresponding terahertz electric field E_{Det} as a function of delay time after correction for the complex-valued detector response of the ZnTe crystal displayed in **b**. **b**, Absolute value (black solid curve) and phase (red dashed curve) of the transfer function[34] calculated for a 6.5-μm-thick ZnTe electro-optic detector in the multi-terahertz frequency range.



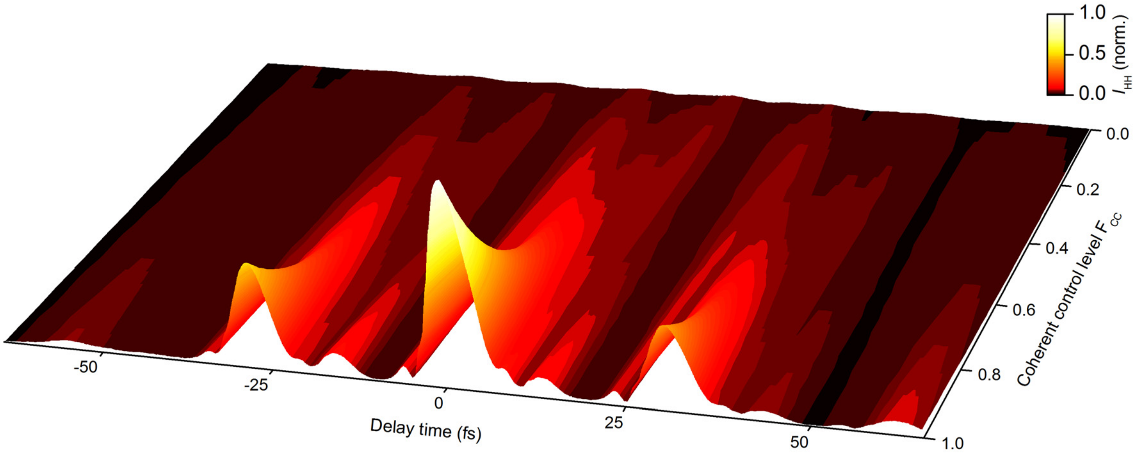
Extended Data Figure 3 | Double-blind reconstruction of the experimental XFROG data. **a**, Retrieved spectral intensities of the detected HH pulse sequence (grey, numerals indicate harmonic orders), the gating pulse (blue) and the sum-frequency intensity (SF) for optimal temporal overlap of gating and HH pulses (black dashed). All spectra displayed here correspond to the data set of main text Fig. 1. The measured spectral intensity of the sum-frequency signal is shown as a red shaded area for comparison. All spectra are normalized to their individual maximum and shifted in intensity for clarity. **b**, Temporal intensity profiles of the gating pulse as measured and reconstructed via SHG-FROG (blue) and double-blind XFROG (black dashed).



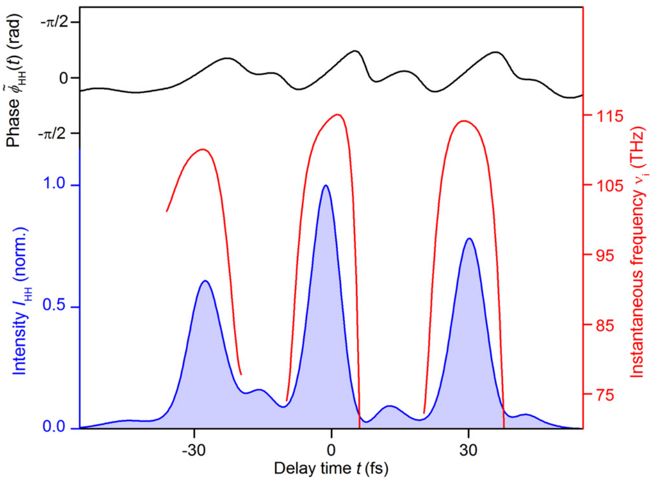
Extended Data Figure 4 | Microscopic model. **a**, Tight binding band structure of GaSe with three valence bands (red) and two conduction bands (blue). The bandgap E_G is marked by the black arrow. **b**, Schematics of interband transitions. The external electric field (orange curve) stimulates a transition between electronic states (blue spheres) in different bands via the dipole moment $d_{h_1e_1}$, indicated by the black arrow. **c**, Schematics depicting intraband currents. Carriers (blue spheres) are accelerated by the external electric field (orange curve) inside their respective bands.



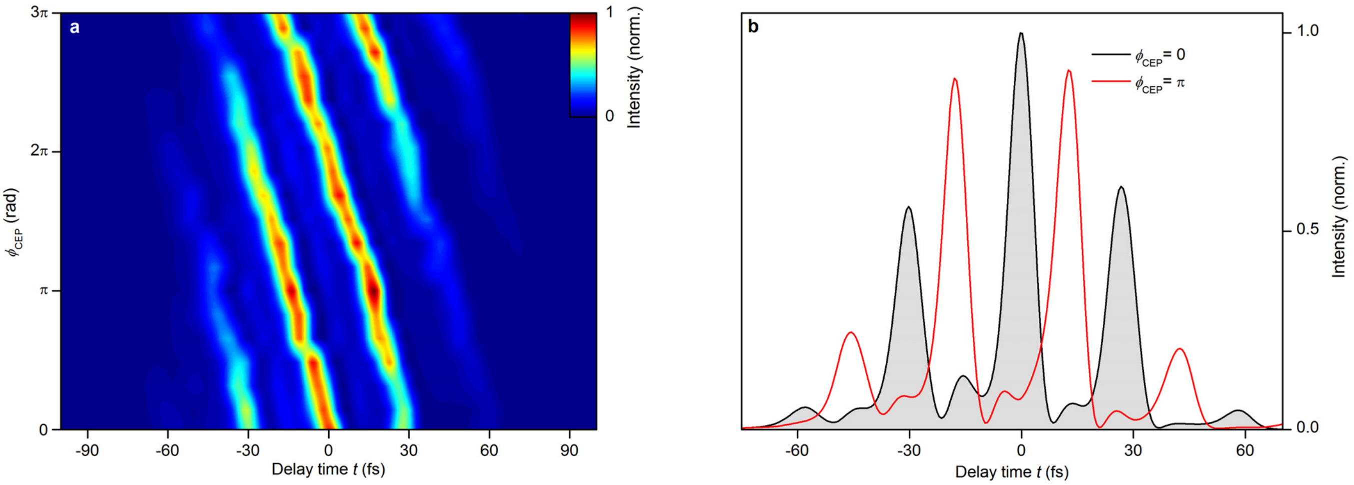
Extended Data Figure 5 | Quantum interference paths. **a**, Simplified band-structure schematics consisting of two valence bands h_1 and h_2 (red lines) and one conduction band e (blue line). The allowed transitions are labelled and marked by black arrows. **b**, Schematic of the direct excitation path. The transition $h_1 \rightarrow e$ is marked by the blue arrow. **c**, Indirect excitation path consisting of the transition $h_1 \rightarrow h_2$ (red arrow) and $h_2 \rightarrow e$ (three-dimensional red arrow). **d**, Full quantum interference (QI): combination of the direct excitation path $h_1 \rightarrow e$ (blue arrow) and the indirect excitation path $h_1 \rightarrow h_2 \rightarrow e$ (red arrows). **e**, The interference path efficiency $\eta_{\mathrm{IPE}}(E_{\mathrm{THz}})$, black line, describes the fraction of additional carriers promoted to the conduction band e_1 by indirect paths such as $h_1 \rightarrow h_2 \rightarrow e$ as a function of the driving terahertz field strength. For the peak fields used in our calculations (indicated by dashed red lines), the strength of indirect excitations is of the same order of magnitude as the strength of direct excitations ($\eta_{\mathrm{IPE}} \propto \mathcal{O}(1)$).



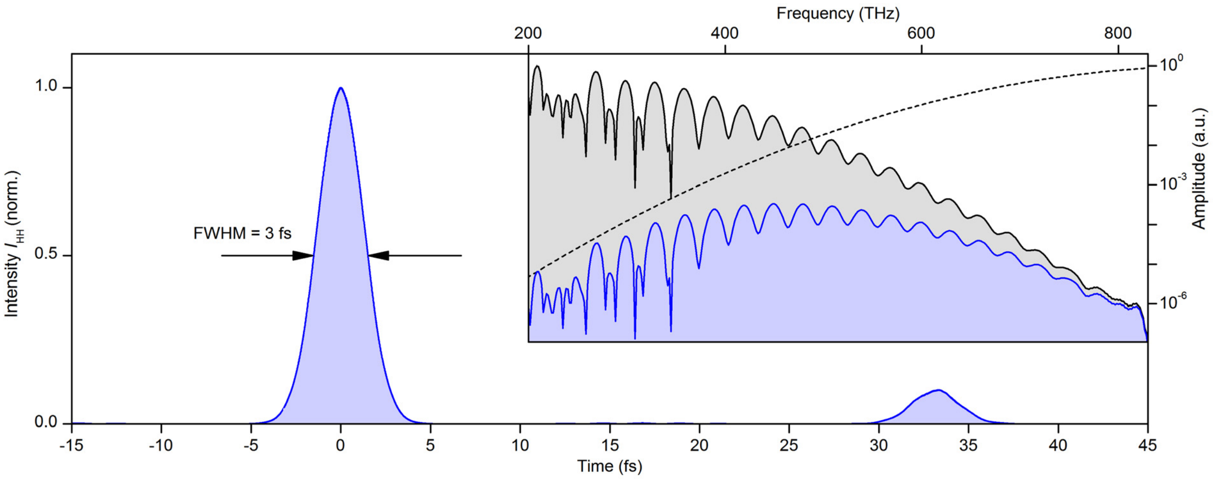
Extended Data Figure 6 | Non-perturbative quantum interference. Globally normalized HH intensity envelopes computed within the five-band model as a function of delay time and the coherent control factor F_{cc} regulating coherent transitions between occupied valence bands. The same data are presented in Fig. 2c in normalized form. Bright colours mark strong emission, dark colours mark weaker emission (colour key at top right).



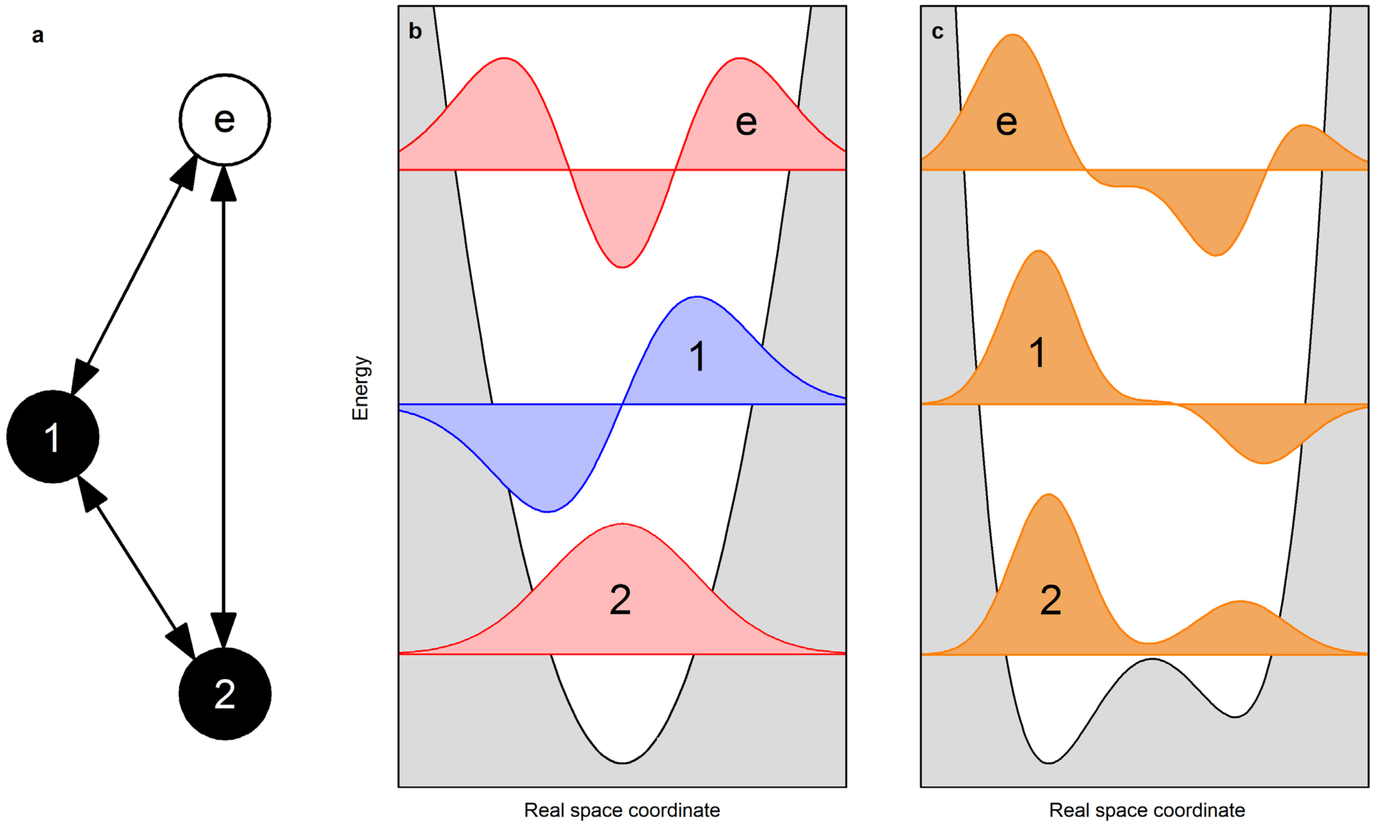
Extended Data Figure 7 | Temporal phase of the HH emission from GaSe. The black curve depicts the retrieved relative phase of typical HH bursts for driving peak fields of 33 MV cm^{-1}. A trivial linear contribution to the phase representing the weighted central frequency ν_c of detected HHs within the detection bandwidth has been subtracted for better visibility: $\tilde{\phi}_{HH}(t) = \phi_{HH}(t) - 2\pi\nu_c t$. The intensity envelope I_{HH} (blue shaded area) and the derived instantaneous frequency ν_i of the three main HH pulses (red curves) are shown on the same timescale.



Extended Data Figure 8 | HH pulse shaping via the CEP of the driving waveform. **a**, False-colour plot of the reconstructed intensity envelopes of the HH pulse sequence versus delay time t for different carrier-envelope phases ϕ_{CEP} of the driving waveform. **b**, Reconstructed HH intensity envelopes for $\phi_{\mathrm{CEP}} = 0$ (black) and $\phi_{\mathrm{CEP}} = \pi$ (red).



Extended Data Figure 9 | Ultrashort pulse shaping via spectral amplitude filtering. Intensity envelope (blue shaded area) of spectrally filtered HH bursts (inset, blue shaded spectrum) emitted from GaSe featuring a full-width at half-maximum of 3 fs only, as calculated with a semiclassical intraband model. Inset, calculated amplitude spectrum driven by a multi-terahertz transient with external peak fields of 43 MV cm^{-1} as emitted from the sample (grey shaded) and spectrally filtered (blue shaded) with a suitable high-pass filter (black dashed).



Extended Data Figure 10 | Triangle system and symmetry. a, Schematic of a triangle system of electronic states as a minimal condition for non-perturbative quantum interference. Filled circles 1 and 2 denote occupied states, whereas the sphere e symbolises an empty state. **b**, A symmetric potential (black solid line) and the resulting wave functions with even (red shaded area) and odd (blue shaded area) parity. **c**, System with an asymmetric potential (black line) resulting in wave functions with indefinite parity (orange shaded areas). See Methods for details.

doi:10.1038/nature14647

Flexible high-temperature dielectric materials from polymer nanocomposites

Qi Li[1], Lei Chen[1], Matthew R. Gadinski[1], Shihai Zhang[2], Guangzu Zhang[1], Haoyu Li[3], Aman Haque[4], Long-Qing Chen[1], Tom Jackson[3] & Qing Wang[1]

Dielectric materials, which store energy electrostatically, are ubiquitous in advanced electronics and electric power systems[1–8]. Compared to their ceramic counterparts, polymer dielectrics have higher breakdown strengths and greater reliability[1–3,9], are scalable, lightweight and can be shaped into intricate configurations, and are therefore an ideal choice for many power electronics, power conditioning, and pulsed power applications[1,9,10]. However, polymer dielectrics are limited to relatively low working temperatures, and thus fail to meet the rising demand for electricity under the extreme conditions present in applications such as hybrid and electric vehicles, aerospace power electronics, and underground oil and gas exploration[11–13]. Here we describe crosslinked polymer nanocomposites that contain boron nitride nanosheets, the dielectric properties of which are stable over a broad temperature and frequency range. The nanocomposites have outstanding high-voltage capacitive energy storage capabilities at record temperatures (a Weibull breakdown strength of 403 megavolts per metre and a discharged energy density of 1.8 joules per cubic centimetre at 250 degrees Celsius). Their electrical conduction is several orders of magnitude lower than that of existing polymers and their high operating temperatures are attributed to greatly improved thermal conductivity, owing to the presence of the boron nitride nanosheets, which improve heat dissipation compared to pristine polymers (which are inherently susceptible to thermal runaway). Moreover, the polymer nanocomposites are lightweight, photopatternable and mechanically flexible, and have been demonstrated to preserve excellent dielectric and capacitive performance after intensive bending cycles. These findings enable broader applications of organic materials in high-temperature electronics and energy storage devices.

The best commercially available dielectric polymer represented by biaxially oriented polypropylene (BOPP) can operate only at temperatures below 105 °C (ref. 14). Therefore, thermal management is always required to enable the use of dielectric polymers in high-temperature applications. For example, to accommodate BOPP film capacitors in the power inverters of hybrid and electric vehicles, which are used to control and convert direct current from batteries into the alternating current required to power the motor, cooling systems have to be employed to decrease the environmental temperature from about 140 °C to about 70 °C. This brings extra weight, volume and energy consumption to the integrated power system and reduces its reliability and efficiency. The upsurge in lightweight and flexible electronic devices has also created a tremendous demand for high-temperature dielectric polymers, as the heat generated by electronic devices and circuitry increases exponentially with miniaturization and functionality.

A variety of high-performance engineering polymers have been considered as possible high-temperature dielectric materials to address these urgent needs[15–19]. Until now, the key criteria established for evaluating high-temperature dielectric polymers has been the glass transition temperature (T_g) and thermal stability. At temperatures approaching T_g, polymers lose their dimensional and electromechanical stability and display large variations in dielectric constant (K) and dissipation factor (DF) with temperature. Making use only of materials that have high T_g works reasonably well at high temperatures but only under relatively low electric fields and this approach has had very limited success when the material is subject to both high temperatures and high voltages.

Here (see Fig. 1a) we thermally crosslinked divinyltetramethyldisiloxane-bis(benzocyclobutene) (BCB) in the presence of boron nitride nanosheets (BNNSs, Fig. 1b and c) to afford the crosslinked nanocomposite *c*-BCB/BNNS (Fig. 1f and g, Supplementary Information section 1). BNNSs, which form a wide-bandgap (~6 eV) insulator with ultrahigh thermal conductivities in the range ~300–2,000 W m^{-1} K^{-1} (refs 20 and 21), were prepared through liquid-phase exfoliation of hexagonal boron nitride (*h*-BN) powders[22].

Compared to the crosslinked pristine BCB referred to as *c*-BCB, the most striking feature of *c*-BCB/BNNS is substantially suppressed high-field electrical conduction at high temperatures (Supplementary Information sections 2 and 3). For example, the electrical conductivity decreases from 4×10^{-12} S m^{-1} in *c*-BCB to 9.2×10^{-14} S m^{-1} in *c*-BCB/BNNS and the conduction loss decreases from 18% in *c*-BCB to 3% in *c*-BCB/BNNS under an applied field of 200 MV m^{-1} at 150 °C. Coupled with the higher Young's modulus arising from the introduced BNNSs, which impedes the occurrence of the electromechanical breakdown[23], the largely reduced electrical conduction in *c*-BCB/BNNS results in a greatly improved Weibull breakdown strength (E_b) at high temperatures. For example, E_b improves from 262 MV m^{-1} for *c*-BCB to 403 MV m^{-1} for *c*-BCB/BNNS with 10 vol% BNNSs at 250 °C. We note that polymer nanocomposites were previously designed towards improved capacitive energy storage, including BNNS-containing nanocomposites[24], and were mainly intended for room temperature applications[25]. For example, at $E_b = 200$ MV m^{-1}, the dielectric loss of the best ferroelectric polymer nanocomposite with BNNSs is 15% at room temperature[24], but this quickly rises to 76% when the temperature is increased to 70 °C (Supplementary Fig. 25).

The dielectric properties of *c*-BCB/BNNS have been evaluated along with state-of-the-art high-temperature capacitor-grade polymer films including polycarbonate (PC, $T_g \approx 150$ °C), poly(ether ether ketone) (PEEK, $T_g \approx 150$ °C), polyetherimide (PEI, $T_g \approx 217$ °C), fluorene polyester (FPE, $T_g \approx 330$ °C) and polyimide (Kapton PI (from Dupont), $T_g \approx 360$ °C) (Supplementary Table 1). We first examine K and DF as a function of temperature and frequency (Fig. 2, Supplementary Information section 4). At 10^4 Hz, which is the frequency of interest for common power conditioning, a minor variation in K with temperature, that is, <1.7%, is seen in *c*-BCB/BNNS from room temperature to 300 °C, while FPE, the next-best dielectric investigated in this study, shows a K variation of over 8% at 300 °C

[1]Department of Materials Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. [2]PolyK Technologies, State College, Pennsylvania 16802, USA. [3]Department of Electrical Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. [4]Department of Mechanical and Nuclear Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA.

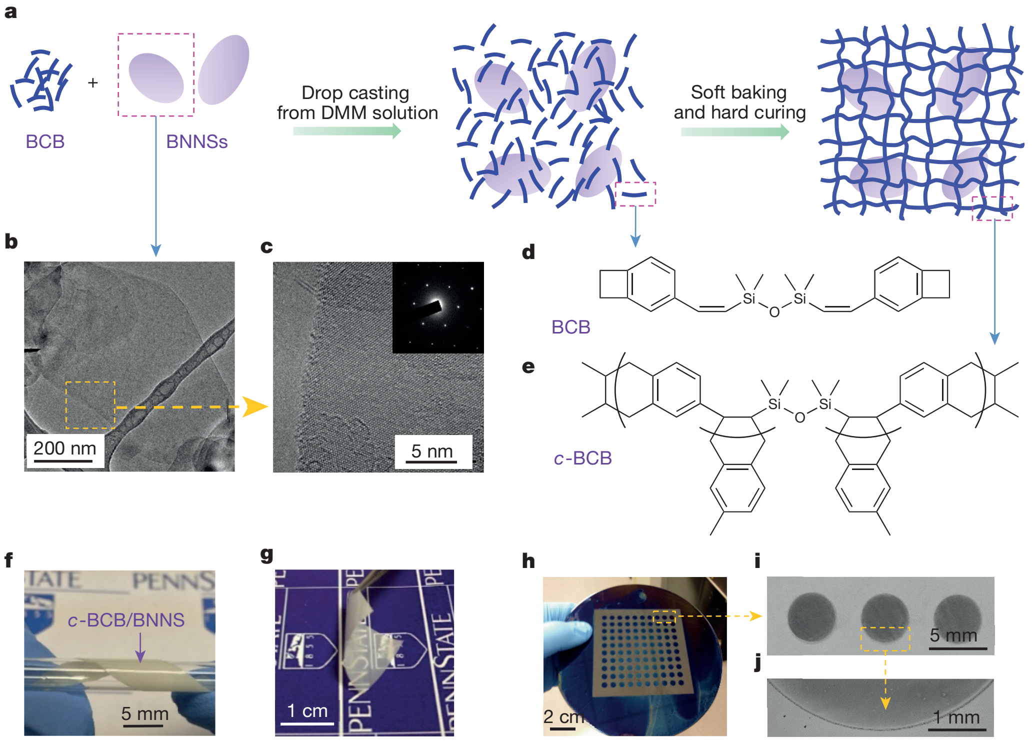


Figure 1 | Material preparation and structures. a, Schematic of the preparation of *c*-BCB/BNNS films. **b**, **c**, Transmission electron microscopy (TEM) images of BNNSs exfoliated from *h*-BN powders. Inset to **c** is an electron-diffraction pattern of BNNSs, showing its hexagonal symmetry. **d**, Chemical structure of the BCB monomer. **e**, The repeating unit of *c*-BCB. **f**, Photographs of a 10-μm-thick *c*-BCB/BNNS film wrapped around a glass tube with diameter 4 mm. **g**, A bent 10-μm-thick *c*-BCB/BNNS film. **h**, The photopatterned *c*-BCB/BNNS on a Si wafer. **i**, **j**, Optical microscopic images of the patterned films; the dark regions correspond to *c*-BCB/BNNS.

relative to room temperature (Supplementary Fig. 15). As presented in Fig. 2c, the temperature coefficient of *K* for *c*-BCB/BNNS is around 65 parts per million (p.p.m.) per °C, compared to 308 p.p.m. $^{\circ}\mathrm{C}^{-1}$ and 498 p.p.m. $^{\circ}\mathrm{C}^{-1}$ for FPE and Kapton, respectively, within the temperature range 25–300 °C. Even under a direct-current bias voltage of 50 MV m^{-1}, the *K* variation of *c*-BCB/BNNS is still as low as 1.6% at 250 °C, compared to 8.5% for FPE (Supplementary Fig. 15).

Concurrently, the DF value of *c*-BCB/BNNS at 10^4 Hz only increases from 0.09% to 0.13% with increasing temperature up to 300 °C (Supplementary Fig. 15). Although Kapton shows a stability of DF with temperature under direct-current bias similar to that of *c*-BCB/BNNS, appreciable increases in DF have been observed in all the other polymer dielectrics. For example, the DF of FPE jumps from 0.22% at room temperature to 1.35% at 280 °C. It is also evident that, of

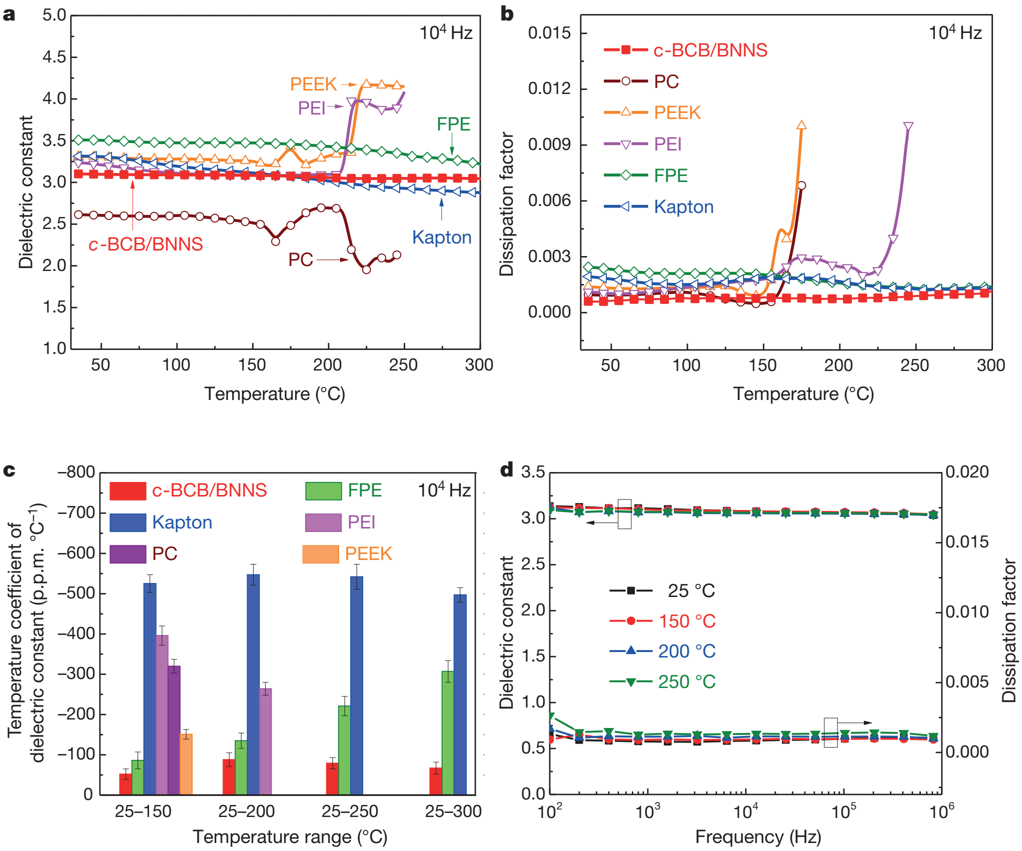


Figure 2 | Dielectric stability. a, Temperature dependence of dielectric constant. **b**, The DF of *c*-BCB/BNNSs with 10 vol% of BNNSs and high-T_g polymer dielectrics. **c**, Temperature coefficient of the dielectric constant of *c*-BCB/BNNS with 10 vol% of BNNSs and high-T_g polymer dielectrics at various temperature ranges. **d**, Frequency dependence of the dielectric constant and DF of *c*-BCB/BNNS with 10 vol% of BNNSs at different temperatures. Error bars show standard deviation.

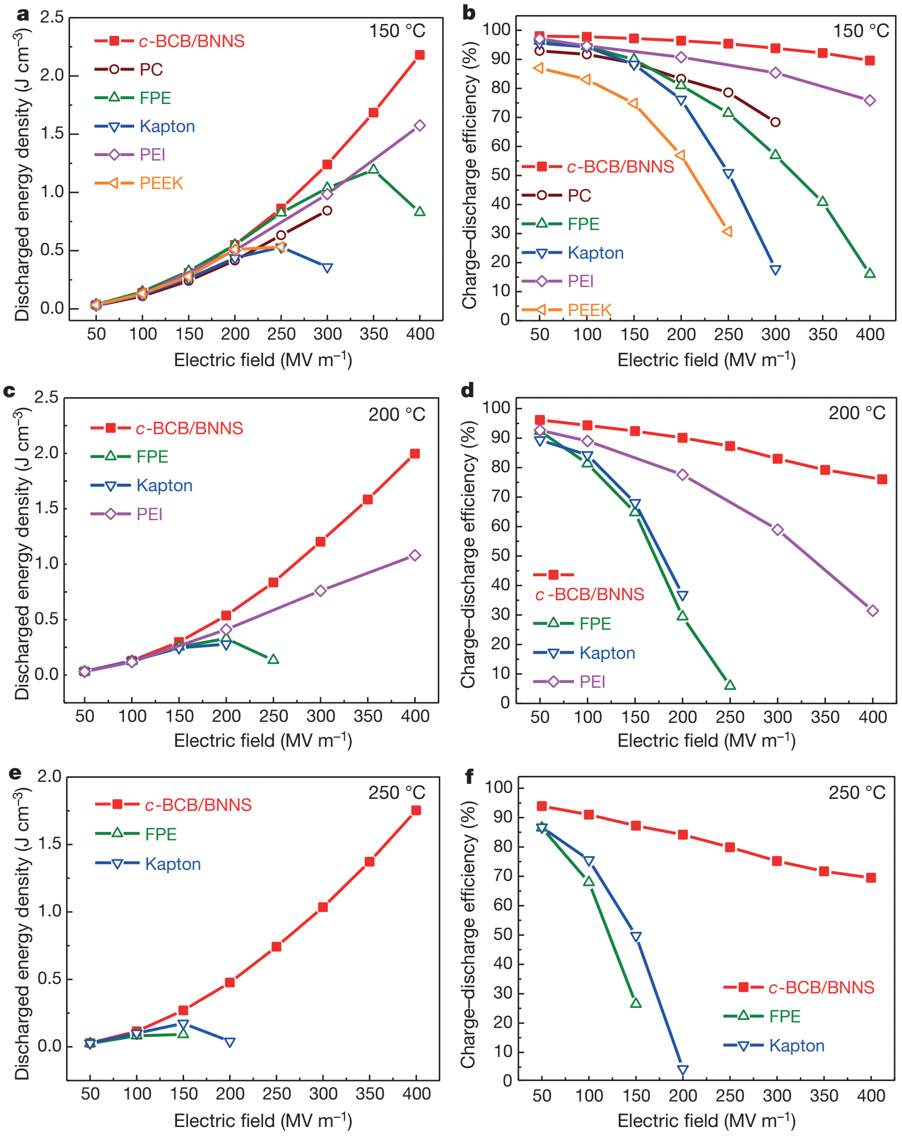


Figure 3 | Electrical energy storage capability. Discharged energy density and charge–discharge efficiency of *c*-BCB/BNNS with 10 vol% of BNNSs and high-T_g polymer dielectrics measured at 150 °C (**a**, **b**), 200 °C (**c**, **d**) and 250 °C (**e**, **f**).

the dielectrics assessed, *c*-BCB/BNNS offers the most stable K and DF in the frequency range 10^2–10^6 Hz at high temperatures (Fig. 2d, Supplementary Figs 16 and 17).

We next studied high-field capacitive energy storage properties at high temperatures (Supplementary Information section 5). As summarized in Fig. 3, *c*-BCB/BNNS clearly outperforms all the high-T_g polymer dielectrics at temperatures ranging from 150 °C to 250 °C in terms of the discharged energy density (U_e) and the charge–discharge efficiency (η). For example, *c*-BCB/BNNS can discharge a U_e exceeding 2.2 J cm^{-3} under 400 MV m^{-1} with a η of larger than 90% at 150 °C. At 200 °C, *c*-BCB/BNNS delivers a U_e of 2 J cm^{-3} under 400 MV m^{-1}, which is twice that of PEI, accompanied by a η value more than 1.5 times higher than that of PEI. As the temperature is further raised to 250 °C, where none of the high-T_g polymer dielectrics can operate at more than 150 MV m^{-1}, *c*-BCB/BNNS is functional up to 400 MV m^{-1} with a U_e of ~1.8 J cm^{-3}. Remarkably, at 200 MV m^{-1}, which is the operating condition of BOPP film capacitors in electric vehicles[26], the value of η for *c*-BCB/BNNS at 150 °C—that is, ~97%—is the same as that of BOPP at 70 °C (Supplementary Fig. 24). This indicates that, by replacing BOPP with *c*-BCB/BNNS, the complex cooling system for power inverters in electric vehicles could be eliminated. Furthermore, under these conditions, the U_e of *c*-BCB/BNNS is over 40% higher than that of BOPP owing to its higher K; that is, 3.1, versus 2.2 for BOPP.

The superior performance of *c*-BCB/BNNS over the high-T_g polymer dielectrics stems from its substantially reduced high-field leakage current at elevated temperatures. For example, at 200 °C and a field of 200 MV m^{-1}, a current density of 4.2×10^{-8} A cm^{-2} is found in *c*-BCB/BNNS, which is nearly one order of magnitude lower than that of PEI and two orders of magnitude smaller than those of FPE and Kapton (Supplementary Fig. 28). It is important to note that the electrical conduction not only accounts for dielectric loss, which degrades U_e and η, but also generates Joule heating within dielectrics[27]. Depending on the heat dissipation (which is determined primarily by the thermal conductivity of the dielectrics), the geometry of the capacitors, and the cooling systems, the steady-state internal temperature of capacitors could exceed the T_g or even the decomposition temperature of dielectric polymers and cause capacitor failure.

We simulate the steady-state internal temperature distribution of the dielectric films by using finite element computations[28] (Methods, Supplementary Information section 6). As summarized in Supplementary Tables 4–6, it can be seen that the steady-state internal temperatures of *c*-BCB/BNNS film are consistently much lower than those of the high-T_g polymers operating under the same conditions, owing to a pronounced reduction in conduction loss and a marked enhancement in thermal conductivity, that is, from ~0.2 W m^{-1} K^{-1} for the polymers to 1.8 W m^{-1} K^{-1} for *c*-BCB/BNNS. As exemplified in Fig. 4, under a forced convection with air (convective heat transfer coefficient $h = 35$) and an ambient temperature of 200 °C, the PEI-, FPE- and Kapton-based film capacitors are overheated, with temperatures at the film centres of 219 °C, 361 °C, and 435 °C, respectively, at 200 MV m^{-1}.

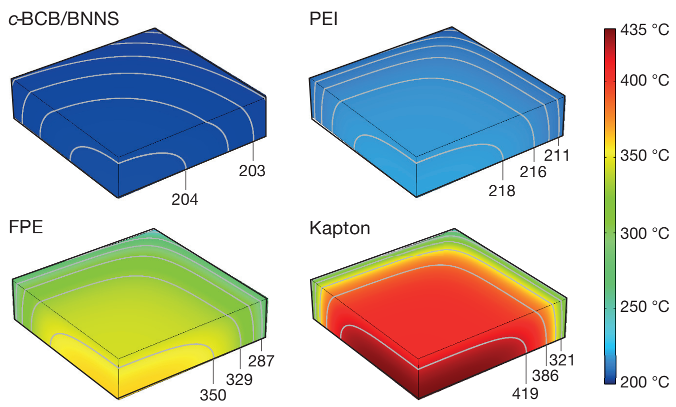


Figure 4 | Steady-state temperature distribution. Simulated steady-state temperature distribution in spiral-wound film capacitors of 40 × 40 × 10 mm^3 enclosure size, based on *c*-BCB/BNNS with 10 vol% of BNNSs, PEI, FPE and PI (Kapton), respectively, given continuous operation under an electric field of 200 MV m^{-1}, with an ambient temperature of 200 °C and a convective heat transfer coefficient of 35. As the model is symmetric in three Cartesian coordinates, only 1/8 of the volume is shown, to expose the centre part of the capacitor.

In contrast, the highest temperature inside *c*-BCB/BNNS is only 204 °C. With the field increasing to 300 MV m^{-1} and 400 MV m^{-1}, the temperatures at the film centre of *c*-BCB/BNNS are 213 °C and 255 °C, respectively, still below its T_g (>350 °C), when $h = 200$. Notably, even at 250 °C and with active liquid cooling, *c*-BCB/BNNS is operable at 300 MV m^{-1} with a maximum internal temperature of 308 °C. Under high-voltage cycling conditions, it is in fact thermal runaway[29] that dictates the maximum operation field of dielectric polymers and thus acts as the limiting factor of high-field capacitive energy storage at elevated temperatures. Despite high T_g and excellent thermal stability from engineering polymers, their well documented poor thermal conductivities[30] seriously limit the actual working temperatures of capacitors under high fields.

Finally, we demonstrate that *c*-BCB/BNNS can be readily prepared by photo-polymerization of solution-cast films, which, upon further curing, are found to possess essentially the same dielectric properties as the thermally crosslinked films (see Methods and Supplementary Fig. 27). The ultraviolet-induced crosslinking through an optical mask enables direct photopatterning of *c*-BCB/BNNS films (Figs 1h–j); this is highly desirable in device fabrications. Furthermore, no degradation in dielectric stability, U_e and η of *c*-BCB/BNNS measured at room temperature and 250 °C was observed after rigorous winding and bending tests (Supplementary Information section 7 and Supplementary Video). This suggests that this nanocomposite may be used in practical flexible electronics and high-throughput roll-to-roll processing into wound cells. Also, it is noteworthy that *c*-BCB/BNNS has the lowest mass density (~1.10 g cm^{-3}) of the polymer dielectrics studied (Supplementary Table 1). This desirable combination of processibility, flexibility, light weight, and dielectric and capacitive performance in such nanocomposites may transform the way compact power modules and power circuits targeted for harsh environment applications are built.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 April; accepted 27 May 2015.

1. Sarjeant, W. J., Zirnheld, J. & MacDougall, F. W. Capacitors. *IEEE Trans. Plasma Sci.* **26,** 1368–1392 (1998).
2. Sarjeant, W. J., Clelland, I. W. & Price, R. A. Capacitive components for power electronics. *Proc. IEEE* **89,** 846–855 (2001).
3. Tan, Q., Irwin, P. & Cao, Y. Advanced dielectrics for capacitors. *IEEJ Trans. Fund. Mater.* **126,** 1152–1159 (2006).
4. Irvine, J. T. S., Sinclair, D. C. & West, A. R. Electroceramics: characterization by impedance spectroscopy. *Adv. Mater.* **2,** 132–138 (1990).
5. Reaney, I. M. & Iddles, D. Microwave dielectric ceramics for resonators and filters in mobile phone networks. *J. Am. Ceram. Soc.* **89,** 2063–2072 (2006).
6. Bell, A. J. Ferroelectrics: the role of ceramic science and engineering. *J. Eur. Ceram. Soc.* **28,** 1307–1317 (2008).
7. Ogihara, H., Randall, C. A. & Trolier-McKinstry, S. High-energy density capacitors utilizing 0.7 $BaTiO_3$–0.3 $BiScO_3$ ceramics. *J. Am. Ceram. Soc.* **92,** 1719–1724 (2009).
8. Xiong, B., Hao, H., Zhang, S. J., Liu, H. X. & Cao, M. H. Structure, dielectric properties and temperature stability of $BaTiO_3$–$Bi(Mg_{1/2}Ti_{1/2})O_3$ perovskite solid solutions. *J. Am. Ceram. Soc.* **94,** 3412–3417 (2011).
9. Chu, B. J. *et al.* A dielectric polymer with high electric energy density and fast discharge speed. *Science* **313,** 334–336 (2006).
10. Ho, J., Jow, T. R. & Boggs, S. Historical introduction to capacitor technology. *IEEE Electr. Insul. Mag.* **26,** 20–25 (2010).
11. Johnson, R. W., Evans, J. L., Jacobsen, P., Thompson, J. R. & Christopher, M. The changing automotive environment: high-temperature electronics. *IEEE Trans. Electron. Packag. Manuf.* **27,** 164–176 (2004).
12. Watson, J. & Castro, G. High-temperature electronics pose design and reliability challenges. *Analog. Dialog* **46,** 1–7 (2012).
13. Weimer, J. A. Electrical power technology for the more electric aircraft. In *Proc. AIAA/IEEE Digital Avionics Systems Conf.* http://dx.doi.org/10.1109/DASC.1993.283509 (IEEE, 1993).
14. Rabuffi, M. & Picci, G. Status quo and future prospects for metallized polypropylene energy storage capacitors. *IEEE Trans. Plasma Sci.* **30,** 1939–1942 (2002).
15. Wang, D. H., Kurish, B. A., Treufeld, I., Zhu, L. & Tan, L. S. Synthesis and characterization of high nitrile content polyimides as dielectric films for electrical energy storage. *J. Polym. Sci. A* **53,** 422–436 (2015).
16. Ho, J. & Jow, T. R. High field conduction in heat resistant polymers at elevated temperature for metallized film capacitors. In *Power Modulator and High Voltage Conf.* http://dx.doi.org/10.1109/IPMHVC.2012.6518764 (IEEE, 2012).
17. Venkat, N. *et al.* High temperature polymer film dielectrics for aerospace power conditioning capacitor applications. *Mater. Sci. Eng. B* **168,** 16–21 (2010).
18. Tan, D., Zhang, L. L., Chen, Q. & Irwin, P. High-temperature capacitor polymer films. *J. Electron. Mater.* **43,** 4569–4575 (2014).
19. Pan, J. L., Li, K., Chuayprakong, S., Hsu, T. & Wang, Q. High-temperature poly(phthalazinone ether ketone) thin films for dielectric energy storage. *ACS Appl. Mater. Interf.* **2,** 1286–1289 (2010).
20. Dean, C. R. *et al.* Boron nitride substrates for high-quality graphene electronics. *Nature Nanotechnol.* **5,** 722–726 (2010).
21. Sevik, C., Kinaci, A., Haskins, J. B. & Çağın, T. Characterization of thermal transport in low-dimensional boron nitride nanostructures. *Phys. Rev. B* **84,** 085409 (2011).
22. Coleman, J. N. *et al.* Two-dimensional nanosheets produced by liquid exfoliation of layered materials. *Science* **331,** 568–571 (2011).
23. Ieda, M. Dielectric breakdown process of polymers. *IEEE Trans. Electr. Insul.* **15,** 206–224 (1980).
24. Li, Q. *et al.* Solution-processed ferroelectric terpolymer nanocomposites with high breakdown strength and energy density utilizing boron nitride nanosheets. *Energy Environ. Sci.* **8,** 922–931 (2015).
25. Dang, Z. M., Yuan, J., Yao, S. H. & Liao, R. J. Flexible nanodielectric materials with high permittivity for power energy storage. *Adv. Mater.* **25,** 6334–6365 (2013).
26. Montanari, D. *et al.* Film capacitors for automotive and industrial applications. In *Proc. CARTS USA* 23–38 (Electronic Components Industry Association, 2009).
27. O'Dwyer, J. J. *The Theory of Electrical Conduction and Breakdown in Solid Dielectrics* Ch. 1 (Clarendon, 1973).
28. Qin, S., Ho, J., Rabuffi, M., Borelli, G. & Jow, T. R. Implications of the anisotropic thermal conductivity of capacitor windings. *IEEE Electr. Insul. Mag.* **27,** 7–13 (2011).
29. Zebouchi, N. *et al.* Electrical breakdown theories applied to polyethylene terephthalate films under the combined effects of pressure and temperature. *J. Appl. Phys.* **79,** 2497–2501 (1996).
30. Mark, J. E. *Physical Properties of Polymers Handbook* Ch.10 (AIP Press, 1996).

Supplementary Information is available in the online version of the paper.

Acknowledgements Q.W. acknowledges financial support from the US Office of Naval Research under grant number N00014-11-1-0342. L.-Q.C. is supported by the Air Force Office of Scientific Research under grant number FA9550-14-1-0264. H.L. and T.J. acknowledge support from the Dow Chemical Corporation.

Author Contributions Q.W. and Q.L. devised the original concept. Q.L. and M.R.G. were responsible for materials synthesis and characterization. Q.L., M.R.G. and G.Z performed dielectric and polarization-loop measurements. L.C. and L.-Q.C. carried out simulation studies. S.Z. and Q.L. performed the studies of high-T_g dielectric polymers. H.L. and T.J. designed the bending tests. A.H. measured thermal conductivities. Q.W. and Q.L. wrote the first draft of the manuscript, and all authors participated in manuscript revision.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Q.W. (wang@matse.psu.edu).

METHODS

Materials. Dipropylene glycol dimethyl ether (DMM), BCB monomers and *b*-staged BCBs (partially polymerized) with the number average molecular weight of ~25,000 were provided by Dow Chemical. Boron nitride powders were purchased from Sigma-Aldrich. The high-T_g polymer dielectric films were provided by Poly*K* Technologies. The polyimide (Kapton) films were vacuum dried overnight at 70 °C before use. All the other materials were used as received.

Preparation. BNNSs were prepared from boron nitride powders using a solution phase exfoliation method[22]. To make the nanocomposites, BNNSs were first dispersed in DMM at a concentration of 5 mg ml^{-1}. 100 mg of BCB monomers were dissolved in 2 ml of DMM and stirred for 2 h. Afterwards the DMM solution of BNNSs was mixed with BCB solution in proportion, and the mixture was first stirred for 10 min and then sonicated for 5 min using a tip-type sonicator (175 W). To crosslink the material, the mixture was drop-cast on a glass slide and subject to baking at 120 °C for 30 min, which was followed by curing at 250 °C for 2 h under N_2. The film was peeled off after soaking in water for 5 min. The film thickness can be varied by tuning the concentration of the cast solution. The thickness of films used for electrical characterizations is within the range of 6–12 µm. For ultraviolet-induced crosslinking of the material, DMM solution of *b*-staged BCB mixed with BNNSs was used to cast a film on Si wafer by spin-coating. Then the spin-coated film was subject to a 20-min ultraviolet exposure in an ultraviolet crosslinker (XL-1500, Spectroline) equipped with ultraviolet tubes (BLE-1T155, Spectroline). The thickness of the film can be varied by changing the solution concentration and parameters of spin-coating. Typically, a 30 wt% solution spin-coated at 700 r.p.m. yields a 4-µm-thick film, which, after curing at 250 °C for 15 min under N_2, achieves the same level of performance as that of the thermally crosslinked films. For a typical procedure of photopatterning, the material was spin-coated on Si wafer from a 20 wt% solution of *b*-staged BCB mixed with BNNSs at 3,000 r.p.m., and then covered with an optical mask before ultraviolet irradiation. Afterwards, the material was developed using DMM and dried.

Characterization. Fourier-transform infrared (FTIR) spectra were obtained in the attenuated total reflectance (ATR) mode using a ZnSe crystal as a contact to the samples with a Varian Digilab FTS-8010 spectrometer. Differential scanning calorimetry was conducted by using a TA Instrument Q100 differential scanning calorimeter at a heating/cooling rate of 10 °C min^{-1}. X-ray diffraction analysis was studied using a PANalytical X'pert Pro MPD theta-theta diffractometer. Thermogravimetric analysis was performed with a TGA 2050 Analyzer at a heating rate of 10 °C min^{-1}. TEM images were obtained on a JEOL JEM-2001F transmission electron microscope. Scanning electron microscopy measurements were performed with a FEI Nova NanoSEM 630 field emission electron microscope. Gold electrodes of diameter 6 mm and thickness 60 nm were sputtered on both sides of the polymer films for the electrical measurements. Dielectric spectra were acquired over a broad temperature range using a Hewlett Packard 4284A LCR meter in conjunction with a Delta Design oven model 2300. Dielectric spectra under direct current bias were collected with the same equipment along with a Hewlett Packard 4140B pA meter/voltage source, a KEPCO BOP 1000M amplifier and a protective circuit. Conduction currents were obtained under an electric field provided by a Hewlett Packard 4140B pA meter/voltage source and TREK model 2210 amplifier. High-field electric displacement–electric field loops were collected using a modified Sawyer–Tower circuit, where the samples were subject to a triangular unipolar wave with a frequency of 10 Hz. Dielectric breakdown strength measurements were performed on a TREK P0621P instrument using the electrostatic pull-down method under a direct-current voltage ramp of 500 V s^{-1}. Young's moduli were derived from strain–stress curves measured with a TA RSA-G2 Solids Analyzer, using a constant linear stretching rate of 0.02% s^{-1}. See Supplementary Information section 6 for the method of thermal conductivity measurement.

Two-parameter Weibull statistic. Dielectric breakdown strength is analysed within the framework of a two-parameter Weibull statistic described as:

$$P(E) = 1 - \exp(-(E/\alpha)^{\beta})$$

where $P(E)$ is the cumulative probability of electric failure, E is the measured breakdown field, the scale parameter α is the field strength for which there is a 63% probability for the sample to breakdown (Weibull breakdown strength), the shape parameter β evaluates the scatter of data and a higher value of β represents greater dielectric reliability.

Temperature coefficient of dielectric constant. The temperature coefficient of dielectric constant, τ_{ε_r}, for a given temperature range (from T_i to T_f), is defined as:

$$\tau_{\varepsilon_r} = (K_f - K_i)/[K_{ref}(T_f - T_i)]$$

where K_{ref} is the dielectric constant at room temperature, T_i and T_f are the low-end and high-end temperatures, respectively, and K_i and K_f are the dielectric constants at T_i and T_f, respectively.

Simulation of steady-state temperature distribution. The full-field temperature evolution in the metallized thin film capacitor during joule heating is mathematically governed by

$$\rho_m C \frac{\partial T(\boldsymbol{x})}{\partial t} = K\nabla^2 T(\boldsymbol{x}) + \sigma(\boldsymbol{x}, T)E^2 \tag{1}$$

where ρ_m and C are the density and heat capacity respectively. K is the thermal conductivity and E stands for the applied electrical field. In particular, as shown by the experimental measurements, the electrical conductivity σ depends on the temperature T, having the characteristic form of:

$$\sigma = \sigma_0 \exp\left(-\frac{A}{k_B T}\right) \tag{2}$$

where the coefficients σ_0 and A for each material at one specific applied electric field are calculated by fitting the measured data and k_B is the Boltzmann constant. Setting $\partial T(\boldsymbol{x})/\partial t = 0$ yields the steady-state temperature solution of equation (1):

$$K\nabla^2 T(\boldsymbol{x}) + \sigma(\boldsymbol{x}, T)E^2 = 0 \tag{3}$$

The governing equation is then solved by finite element simulations using the commercial software Comsol 5.0 (http://www.comsol.com/release/5.0). When the capacitor is operating the thin slices of the structures are packed into an enclosure, with the cooling liquid surrounding the enclosure. Therefore, the capacitor is treated as an integrated body with a uniform applied electric field for each calculation. The geometries arising from the experimental setup are used in the finite element model. As a result of the laminar structure of the capacitor, in which the polymer has a relatively low thermal conductivity compared to the metallization, the anisotropy in thermal the conductivity is considered in the model (as shown in Supplementary Information section 6). During the finite element simulation, the heat flux at all the enclosure surfaces of the capacitor is produced by both convective and radiative heat transfer arising from the surrounding cooling, with the form:

$$\boldsymbol{n}(-K\nabla T) = h(T - T_{ext}) + \sigma_{SB}\varepsilon(T^4 - T_{ext}^4) \tag{4}$$

where T_{ext} is the surrounding temperature, h is convective heat transfer coefficient, σ_{SB} is the Stefan–Boltzmann constant, and ε is the emissivity of the surfaces. Owing to symmetry only 1/8 of the volume is modelled. Based on the mesh-sensitive study, the number of mesh elements in the system is set as 120 × 120 × 30 with a grid spacing of 0.33 mm.

doi:10.1038/nature14598

Onset of Antarctic Circumpolar Current 30 million years ago as Tasmanian Gateway aligned with westerlies

Howie D. Scher[1], Joanne M. Whittaker[2], Simon E. Williams[3], Jennifer C. Latimer[4], Wendy E. C. Kordesch[5] & Margaret L. Delaney[6]

Earth's mightiest ocean current, the Antarctic Circumpolar Current (ACC), regulates the exchange of heat and carbon between the ocean and the atmosphere[1], and influences vertical ocean structure, deep-water production[2] and the global distribution of nutrients and chemical tracers[3]. The eastward-flowing ACC occupies a unique circumglobal pathway in the Southern Ocean that was enabled by the tectonic opening of key oceanic gateways during the break-up of Gondwana (for example, by the opening of the Tasmanian Gateway, which connects the Indian and Pacific oceans). Although the ACC is a key component of Earth's present and past climate system[1], the timing of the appearance of diagnostic features of the ACC (for example, low zonal gradients in water-mass tracer fields[4–7]) is poorly known and represents a fundamental gap in our understanding of Earth history. Here we show, using geophysically determined positions of continent–ocean boundaries[8], that the deep Tasmanian Gateway opened 33.5 ± 1.5 million years ago (the errors indicate uncertainty in the boundary positions). Following this opening, sediments from Indian and Pacific cores recorded Pacific-type neodymium isotope ratios, revealing deep westward flow equivalent to the present-day Antarctic Slope Current. We observe onset of the ACC at around 30 million years ago, when Southern Ocean neodymium isotopes record a permanent shift to modern Indian–Atlantic ratios. Our reconstructions of ocean circulation show that massive reorganization and homogenization of Southern Ocean water masses coincided with migration of the northern margin of the Tasmanian Gateway into the mid-latitude westerly wind band, which we reconstruct at 64° S, near to the northern margin. Onset of the ACC about 30 million years ago coincided with major changes in global ocean circulation[9] and probably contributed to the lower atmospheric carbon dioxide levels that appear after this time[10].

The development of a deep oceanic passage in the Tasmanian Gateway between Australia and Antarctica has been linked to the onset of the ACC[11]. However, the relationship between the tectonic evolution of the Southern Ocean and the appearance of diagnostic features of the present-day ACC is poorly known. The formation of the Tasmanian Gateway occurred as a result of the continental break-up of Gondwanaland during the late Cretaceous and Palaeogene periods[12]. Australia and Antarctica separated about 83 million years ago (Ma), although seafloor around the Tasmanian Gateway subsided slowly[13], and a seawater connection was delayed until after 49 Ma (ref. 14). The first ocean current in the Tasmanian Gateway was a shallow water current travelling from the Pacific into the Indian Ocean (that is, westward) under the influence of the polar easterly winds, as the gateway occupied a more southerly position in the middle Eocene[14]. Today, the eastward-flowing ACC is driven by the mid-latitude westerly winds, which produce a southerly pressure gradient force that balances wind-driven northward Ekman transport of surface water, resulting in a geostrophic current that extends to the seafloor[15]. These findings indicate that onset of the ACC must have been accompanied by a reversal in the water-mass tracer field through the Tasmanian Gateway. Moreover, the results imply that the position of the gateway relative to the wind field was essential for ACC onset[16]. We test the viability of this hypothesis by reconstructing absolute palaeolatitudes of the gateway through time, the position of the polar front (that is, the boundary between the polar easterly and mid-latitude westerly winds) during the Oligocene epoch, and the evolution of the zonal gradient of neodymium (Nd) isotopes in the Southern Ocean.

We reconstruct the tectonic opening of the Tasmanian Gateway[17] in a moving hotspot reference frame[18], using geophysically determined continent–ocean boundaries (COBs) from the South Tasman Rise and Antarctica[8]. Opening of the Tasmanian Gateway below around 500 m (that is, upper continental slope) is indicated by separation of the COBs (Fig. 1). Defining the innermost and outermost possible COBs—to account for uncertainty in the COB location—produces a range of ages for the final separation of Australia and Antarctica of 35–32 Ma (Fig. 2 and Extended Data Fig. 1). During this interval, the northern boundary of the nascent gateway was positioned to the south of 65° S.

To determine whether the gateway was in the latitude band of the polar easterlies at this time, we reconstruct the palaeowind field by taking the palaeopolar front to be the boundary between the polar easterly and westerly winds. The present boundary between the mid-latitude westerlies and the polar easterlies coincides with a zone of deep-water upwelling just south of the polar front (Fig. 1a). We reconstruct the latitude of the Oligocene polar front by using microfossil assemblages from sediment cores[19], restoring the 30-Ma positions of the cores used to demarcate the front (yellow dots in Fig. 1b; see also Extended Data Table 1). Combining these approaches, our reconstructions confirm that the Tasmanian Gateway was positioned within the polar easterlies during the final separation of Australia from Antarctica (Fig. 2c). The absolute reference frame used does not alter the relative relationship between the northern edge of the gateway and the Oligocene polar front; however, it does have an influence on the absolute latitude of these features (Extended Data Fig. 2).

To assess the response of ocean circulation to the changes in palaeogeography indicated by the reconstruction of tectonic plates, we analysed the ratios of Nd isotopes found in fossil fish teeth from Ocean Drilling Program (ODP) sites 1124 (Hikurangi Plateau), 1168 (Western Tasmanian Margin) and 1172 (East Tasman Plateau); see Extended Data Table 2 for details of the study sites. Fish teeth incorporate seawater Nd during a post-mortem mineralogical transformation on the seafloor; the inherited Nd signal (ε_{Nd}, where ε_{Nd} represents the $^{143}Nd/^{144}Nd$ ratio of a sample relative to that of the bulk Earth, in

[1]Department of Earth and Ocean Sciences, University of South Carolina, Columbia, South Carolina 29208, USA. [2]Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania 7001, Australia. [3]EarthByte group, School of Geosciences, The University of Sydney, Sydney, New South Wales 2006, Australia. [4]Department of Earth and Environmental Systems, Indiana State University, Terre Haute, Indiana 47809, USA. [5]Department of Ocean and Earth Science, National Oceanography Centre, University of Southampton, Waterfront Campus, European Way, Southampton SO14 3ZH, UK. [6]Ocean Sciences Department and Institute of Marine Sciences, University of California Santa Cruz, Santa Cruz, California 95064, USA.

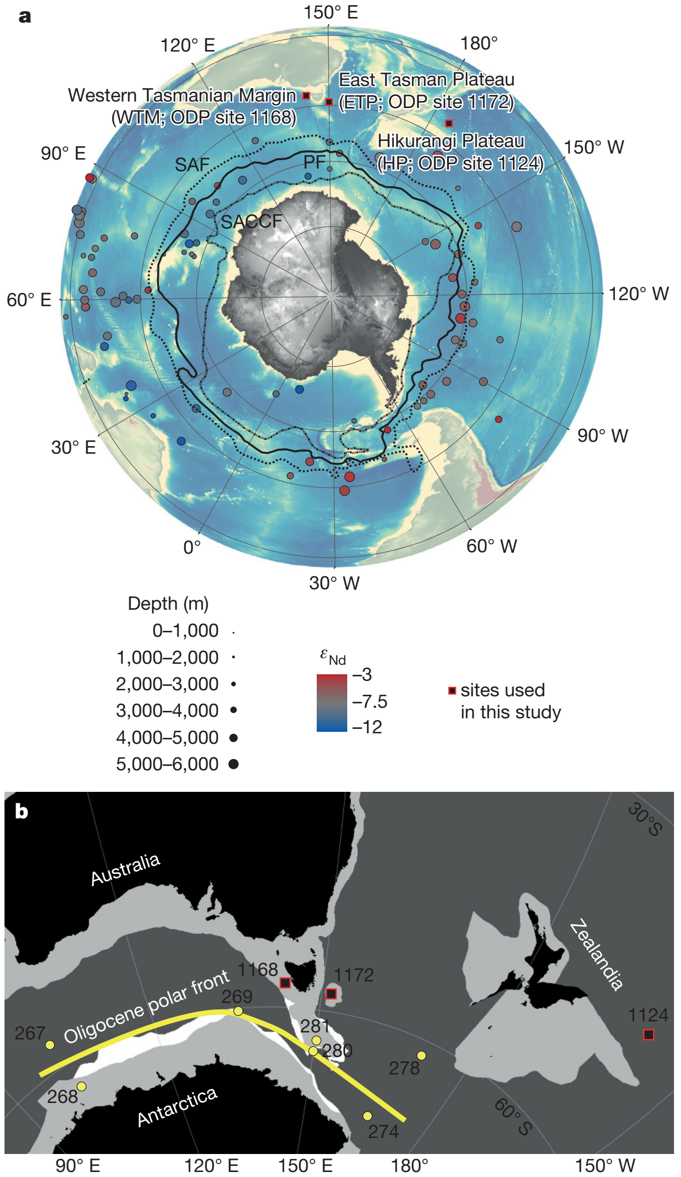


Figure 1 | Maps of the present-day Southern Ocean and relevant study sites. **a**, Map showing study sites, present-day zonal ε_{Nd} distribution, and major frontal zones. Black squares with red borders show the present-day locations of the sediment cores used in this study. Black lines show the meridional extent of the major frontal zones associated with the ACC[30]. Circles show the locations of Southern Ocean ferromanganese (Fe–Mn) nodules found on the seafloor[7]. The colour of the circles shows the ε_{Nd} values of the surface layers; these surface layers are in equilibrium with overlying bottom water. The size of the circles reflects water depth. SAF, Sub-Antarctic Front; PF, Polar Front; SACCF, Southern Antarctic Circumpolar Current Front. **b**, Reconstruction of the early Oligocene (30 Ma) tectonic plates around the Tasmanian Gateway (the narrow gap between Antarctica and Australia). Continents with present-day shorelines are in black. Light grey indicates the continental shelf; dark grey denotes ocean basin/oceanic crust. White bands along the outer continental shelf illustrate the range between the outermost and innermost geophysical expression of the COB on the South Tasman Rise (light grey and white lobe south of Tasmania) and Antarctic conjugate margins[8]. Yellow circles indicate the reconstructed position of Deep Sea Drilling Program (DSDP) sediment cores, with microfossil assemblage data[19] used to reconstruct the Oligocene position of the polar front (yellow band). Red squares indicate the reconstructed positions of Ocean Drilling Program (ODP) sediment cores, used to obtain fossil fish tooth ε_{Nd} records in this study. The plate reconstruction was made using GPlates (http://www.gplates.org).

parts per 10,000), which reveals bottom-water Nd ratios at the time of deposition, is not altered when the teeth become buried in pelagic sediments[20].

Seawater Nd comes mainly from continental crust and has a short residence time in seawater (τ_{Nd} = 300–1,000 years). Thus, water-mass ε_{Nd} is sensitive to water-mass mixing and changes in weathering inputs[21]. Pacific water masses are significantly more radiogenic (that is, have more positive ε_{Nd} values) than are Indian waters (Fig. 2), reflecting the predominance of juvenile mafic crusts that supply Nd to Pacific seawater. The large and continuous interbasin ε_{Nd} gradient between the Pacific and Indian Oceans provides a useful means to monitor the communication of Pacific and Indian water masses through the Tasmanian Gateway (Fig. 2).

The Nd found in fossil fish teeth in our study sites comes from Pacific and Indian deep waters and local terrigenous inputs. Sources of Nd isotope data pertaining to seawater and terrigenous inputs are given in Extended Data Table 3. Around the time of the opening of the Tasmanian Gateway, Pacific deep waters had ε_{Nd} values of about −4, and values in the Indian Ocean were between −6 and −8 (Fig. 2 and Extended Data Fig. 3). Today, the nearest source of terrigenous inputs to the Tasman margins is the Murray River, where dissolved and suspended material has ε_{Nd} values of −6 and −4.7, respectively. Terrigenous inputs will be more influential in the shallow mixed layer (0–400 m in this region[22]), which was unlikely to be deeper than present during the study interval[23].

Nd isotopes are sensitive to water-mass mixing, so it is essential to know the water depth at the time of fish tooth deposition. We use subsidence curves for the Tasman sediment cores[13] (see Methods) to provide constraints on water depth (Fig. 2a). The Western Tasmanian Margin (site 1168) was in shallow water (that is, less than 400 m) until 32 Ma. The ε_{Nd} values at about 36 Ma are within the range of source rocks in southwest Australia (here represented by the Murray River; Fig. 2), indicating that the ε_{Nd} signature of the surface mixed layer was under the influence of nearby terrigenous inputs, consistent with modern observations of Southern Ocean surface waters[24]. From 35 Ma to 32 Ma, ε_{Nd} values point to an enhanced advection of the warm, near-surface proto Leeuwin Current that accompanied early gateway opening[11]. The pathway and ε_{Nd} signature of this current is not well constrained, but, given its origin, an Indian Ocean signature is likely and its influence has been documented near the Australian coast[23]. The East Tasman Plateau (site 1172) was within deep water (more than 2,000 m; see Methods and Extended Data Fig. 4) over the study interval. Fossil fish tooth ε_{Nd} values plot within the Pacific endmember field from around 36 to 30 Ma (where 'endmember' here refers to a chemically distinct water mass that is involved in water-mass mixing). Throughout this period, the Hikurangi Plateau (site 1124) recorded ε_{Nd} values in the range of the Pacific endmember (grey line Fig. 2), consistent with its location and depth (Hikurangi Plateau basement lavas erupted in the early to late Cretaceous[25] and did not experience significant thermal or tectonic subsidence during the study interval).

Once the Western Tasmanian Margin subsided into deeper waters (about 32 Ma; arrow in Fig. 2a), there ceased to be a significant ε_{Nd} gradient between the study sites, indicating that the western Pacific and Indian sectors of the Southern Ocean (below the mixed layer) were influenced by a common water mass. Between 32 and 30 Ma the source of Nd was Pacific seawater, indicating that the first deep-water current to penetrate the Tasmanian Gateway was flowing westward from the Pacific to the Indian Ocean. This current regime affected slope depths on both sides of Tasmania, and was probably equivalent to the present-day Antarctic Slope Current (ASC). The ASC is found at slope depths between 66° S and 64° S in the western Pacific and Indian sectors of the Southern Ocean[26], is the deep-water counterpart to the Antarctic Counter Current, and is driven by the polar easterlies. The Pacific water signature is not observed farther to the west at Kerguelen

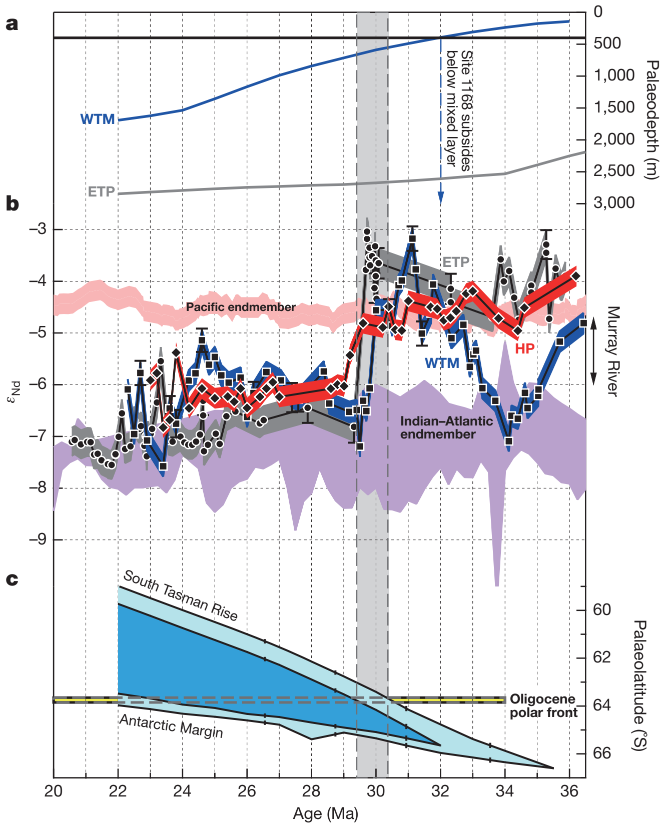


Figure 2 | Subsidence curves, Nd isotope records, and conjugate margin palaeogeography. a, Palaeodepth curves for seafloor on the Western Tasmanian Margin[13] (WTM; site 1168; blue) and East Tasman Plateau (ETP; site 1172; grey). The blue arrow indicates when site 1168 subsided below the mixed layer (black line at 400 m). **b**, Fossil fish tooth Nd isotope records from the WTM, ETP and Hikurangi Plateau (HP; site 1124; red). Error envelopes represent instrumental uncertainty (2σ) based on replicate Nd isotope analyses of the JNdi-1 Nd isotope standard. Error bars (standard error) are shown for samples where the standard error exceeds 2σ instrumental uncertainty. The Pacific and Indian–Atlantic ε_{Nd} envelopes are based on fossil fish tooth Nd isotope records from the equatorial Pacific, southern Kerguelen Plateau, Ninetyeast ridge, and Maud Rise (Extended Data Fig. 3). The range of Murray River dissolved and particulate ε_{Nd} is shown on the right axis. **c**, The palaeolatitudes of the conjugate margins of the Tasmanian Gateway outline the width and position of the gateway during opening. Light blue bands correspond to maximum gateway width, based on geophysically determined positions of the innermost COBs on the South Tasman Rise and Antarctica[8], selected along the narrowest meridional transect of the gateway (Extended Data Fig. 1). Dark blue band shows minimum gateway width based on the outermost COBs. Error bars (standard error) on the boundaries of the gateway reflect uncertainty in the positions of the conjugate margins based on the 95% confidence intervals from the rotations[17]. The Oligocene position of the polar front (yellow band) was derived from microfossil assemblages in deep-sea sediment cores[19], which were incorporated into this plate reconstruction. The grey band centred at 30 Ma highlights when the South Tasman Rise migrated north of the Oligocene polar front.

Plateau or Maud Rise in the early Oligocene (Extended Data Fig. 3), illustrating that different sectors of the Southern Ocean had independent deep-circulation regimes before the onset of the ACC—possibly a deep counterpart to the polar gyres suggested by microfossil distributions and numerical simulations of surface currents[11,27].

Between 30 Ma and 29 Ma, the ε_{Nd} records at the study sites converge with the Indian–Atlantic endmember. This observation indicates that the Pacific sector of the Southern Ocean, at least as far west as the Hikurangi Plateau, came under the influence of a current that flowed eastward from the Indian to the Pacific. The only other ε_{Nd} record of equivalent age in the southwest Pacific (Fe–Mn crust Nova, a hydrogenous deposit dredged from the Nova Canton Trough[28]) shows a similar shift towards Indian–Atlantic seawater ε_{Nd} values. Not only did the direction of deep-water flow reverse at 30 Ma, but the composition of bulk sediment (that is, lithogenous Al/Ti ratios; Extended Data Fig. 5) on the East Tasman Plateau changed, from reflecting mafic sources (typical of Pacific sediments) to suggesting a source akin to average upper continental crust (typical of Indian–Atlantic sediments). The collapse of zonal gradients throughout the Southern Ocean is strong evidence for the onset of the ACC between 30 and 29 Ma.

The collapse of zonal ε_{Nd} gradients in the Southern Ocean coincides with migration of the northern margin of the Tasmanian Gateway into the zone of the westerlies (that is, north of the Oligocene polar front; Fig. 2c). The major reorganization of oceanic circulation between 30 and 29 Ma does not coincide with any significant change in subsidence around the gateway or with the initial opening of a deep gateway. Instead, we propose that northward migration of the gateway into the influence of the westerly winds established the conditions for geostrophic balance in the Southern Ocean that drive the modern ACC. Although our findings do not preclude the influence of far-field tectonics (specifically, the influence of the Drake Passage) on the observed ε_{Nd} patterns, it is clear that the local interaction between the position of the continents and the position of the winds had a strong effect on patterns of Southern Ocean circulation.

General circulation models predict a quantitative link between upwelling in the ACC and the strength of the Atlantic meridional overturning circulation (AMOC)[2,29]. The results of this study confirm this long-held view; collapse of zonal water-mass tracer gradients in the Southern Ocean at 30 Ma coincides with stronger AMOC and a shift towards the modern four-layer ocean structure after 30 Ma (ref. 9). Because the AMOC sequesters carbon into the deep sea as deep water forms in the North Atlantic Ocean, evolution of the modern ocean structure is consistent with an increase in the carbon storage capacity of the deep sea[3]. Thus, the onset of the ACC may help to explain the observed reduction in atmospheric CO_2 after 30 Ma (ref. 10), and ultimately the stabilization of the present-day icehouse climate state.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 January; accepted 27 May 2015.

1. Anderson, R. F. *et al.* Wind-driven upwelling in the Southern Ocean and the deglacial rise in atmospheric CO_2. *Science* **323,** 1443–1448 (2009).
2. Cox, M. D. An idealized model of the world ocean. 1. The global-scale water masses. *J. Phys. Oceanogr.* **19,** 1730–1752 (1989).
3. Toggweiler, J., Russell, J. & Carson, S. Midlatitude westerlies, atmospheric CO_2, and climate change during the ice ages. *Paleoceanography* **21,** PA2005 (2006).
4. Jeandel, C. Concentration and isotopic composition of Nd in the South Atlantic Ocean. *Earth Planet. Sci. Lett.* **117,** 581–591 (1993).
5. Stichel, T., Frank, M., Rickli, J. & Haley, B. The hafnium and neodymium isotope composition of the Atlantic sector of the Southern Ocean. *Earth Planet. Sci. Lett.* **317–318,** 282–294 (2011).
6. Basak, C., Pahnke, K., Frank, M., Lamy, F. & Gersonde, R. Neodymium isotopic characterization of Ross Sea Bottom Water and its advection through the southern South Pacific. *Earth Planet. Sci. Lett.* **419,** 211–221 (2015).
7. Albarède, F., Goldstein, S. L. & Dautel, D. The neodymium isotopic composition of manganese nodules from the Southern and Indian oceans, the global oceanic neodymium budget, and their bearing on deep ocean circulation. *Geochim. Cosmochim. Acta* **61,** 1277–1291 (1997).
8. Williams, S. E., Whittaker, J. M. & Müller, R. D. Full-fit, palinspastic reconstruction of the conjugate Australian-Antarctic margins. *Tectonics* **30,** 21 (2011).
9. Katz, M. E. *et al.* Impact of Antarctic Circumpolar Current development on late Paleogene ocean structure. *Science* **332,** 1076–1079 (2011).
10. Pagani, M. *et al.* The role of carbon dioxide during the onset of Antarctic glaciation. *Science* **334,** 1261–1264 (2011).
11. Stickley, C. E. *et al.* Timing and nature of the deepening of the Tasmanian Gateway. *Paleoceanography* **19,** PA4027 (2004).
12. Cande, S. C. & Mutter, J. C. A revised identification of the oldest sea-floor spreading anomalies between Australia and Antarctica. *Earth Planet. Sci. Lett.* **58,** 151–160 (1982).
13. Hill, P. J. & Exon, N. F. in *The Ceonzoic Southern Ocean: Tectonics, Sedimentation and Climate Change Between Australia and Antarctica* (eds Exon, N. F., Kennet, J. P. & Malone, M.) 19–42 (Geophysical Monograph 151, American Geophysical Union, 2004).

14. Bijl, P. K. *et al.* Eocene cooling linked to early flow across the Tasmanian Gateway. *Proc. Natl Acad. Sci. USA* **110,** 9645–9650 (2013).
15. Gille, S. T. Float observations of the Southern Ocean. Part I: estimating mean fields, bottom velocities, and topographic steering. *J. Phys. Oceanogr.* **33,** 1167–1181 (2003).
16. Hill, D. J. *et al.* Paleogeographic controls on the onset of the Antarctic circumpolar current. *Geophys. Res. Lett.* **40,** 5199–5204 (2013).
17. Cande, S. C. & Stock, J. M. Pacific-Antarctic-Australia motion and the formation of the Macquarie Plate. *Geophys. J. Int.* **157,** 399–414 (2004).
18. O'Neill, C., Müller, D. & Steinberger, B. On the uncertainties in hot spot reconstructions and the significance of moving hot spot reference frames. *Geochem. Geophys. Geosyst.* **6,** Q04003 (2005).
19. Nelson, C. S. & Cooke, P. J. History of oceanic front development in the New Zealand sector of the Southern Ocean during the Cenozoic — a synthesis. *NZ J. Geol. Geophys.* **44,** 535–553 (2001).
20. Martin, E. E. & Scher, H. D. Preservation of seawater Sr and Nd isotopes in fossil fish teeth: bad news and good news. *Earth Planet. Sci. Lett.* **220,** 25–39 (2004).
21. Goldstein, S. L. & Hemming, S. R. in *Treatise on Geochemistry* Vol. 6 (eds Heinrich, H. D. & Turekian, K. K.) 453–489 (Pergamon, 2003).
22. Dong, S., Sprintall, J., Gille, S. T. & Talley, L. Southern Ocean mixed-layer depth from Argo float profiles. *J. Geophys. Res.* **113,** C06013 (2008).
23. Sijp, W. P., England, M. H. & Huber, M. Effect of the deepening of the Tasman Gateway on the global ocean. *Paleoceanography* **28,** 18 (2011).
24. Stichel, T. *et al.* Sources and input mechanisms of hafnium and neodymium in surface waters of the Atlantic sector of the Southern Ocean. *Geochim. Cosmochim. Acta* **94,** 22–37 (2012).
25. Hoernle, K. *et al.* Age and geochemistry of volcanic rocks from the Hikurangi and Manihiki oceanic Plateaus. *Geochim. Cosmochim. Acta* **74,** 7196–7219 (2010).
26. Bindoff, N. L., Rosenberg, M. A. & Warner, M. J. On the circulation and water masses over the Antarctic continental slope and rise between 80 and 150 degrees E. *Deep-Sea Res. II* **47,** 2299–2326 (2000).
27. Huber, M. *et al.* Eocene circulation of the Southern Ocean: was Antarctica kept warm by subtropical waters? *Paleoceanography* **19,** PA4026 (2004).
28. van de Flierdt, T. *et al.* Deep and bottom water export from the Southern Ocean to the Pacific over the past 38 million years. *Paleoceanography* **19,** PA1020 (2004).
29. Toggweiler, J. R. & Samuels, B. Effect of Drake Passage on the global thermohaline circulation. *Deep-Sea Res. I* **42,** 477–500 (1995).
30. Orsi, A. H., Whitworth, T. III & Nowlin, W. D. Jr. On the meridional extent and fronts of the Antarctic Circumpolar Current. *Deep-Sea Res. I* **42,** 641–673 (1995).

Acknowledgements H.D.S. received support from an UTAS IMAS Visiting Scholarship and an EES travel grant while this manuscript was prepared, and was supported by the US National Science Foundation (NSF) award OCE 1155630. J.M.W. was supported by ARC grant DE140100376. M.L.D. was supported by NSF award OCE-0647876. This research used data and samples from the Ocean Drilling Program (ODP), which was sponsored by the NSF and participating countries under the management of Joint Oceanographic Institutions. GPlates is free software that is licensed for distribution under the GNU General Public License v2 (http://www.gplates.org). Researchers in the EarthByte group have made GPlates compatible data available, some of which were used in this study. We thank C. Moy, C. Riesselman, R. Rykaczewski and A. Winguth for discussions. M. Seton and W. Sijp provided feedback on an early version of the manuscript. J. Aggarwal, E. Bair, M. Frank and S. Sampson provided analytical support.

Author Contributions H.D.S. and M.L.D. designed the research; J.M.W. and S.E.W. carried out plate-tectonic modelling and assessed error in the past positions of conjugate margins; H.D.S., J.C.L. and W.E.C.K. collected geochemical data. All authors contributed to interpreting the data and writing the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.D.S. (hscher@geol.sc.edu).

METHODS

Methods of treatment of deep-sea sediments. Fossil fish teeth were hand-picked from the >150-μm size fraction of washed samples. Fossil fish teeth samples consisted of two to five teeth each, ranging in mass from 10 μg to 100 μg Nd. Sample replicates were created in several intervals where fish teeth were abundant. Samples were sonified in quartz-distilled water and methanol to remove debris from surfaces and cavities, then treated with reductive, oxidative, and weak acid steps[31, 32]. Samples were transferred into pre-cleaned microcentrifuge tubes before and after the partial dissolution step. Following final transfer into clean tubes, 50 μl of distilled 0.25 M HCl was added to the microcentrifuge tubes to dissolve the samples before column chemistry.

Nd isotope analyses. All of the samples used in this study were processed through the single column method[33], with the only change being a doubling of the column length to improve separation of samarium. Cation exchange chemistry and Nd isotopic analysis of samples from ODP site 1172 were done in the W. M. Keck Radiogenic Isotope Facility at UC Santa Cruz. Cation exchange and isotopic analysis of samples from ODP sites 1168 and 1124 was done in the Center for Elemental Mass Spectrometry at the University of South Carolina.

All measurements were made on a Neptune multiple collector inductively coupled plasma mass spectrometer (MC-ICP-MS) with an Apex HF or Apex Q as the introduction system. An X skimmer cone was used. All Nd isotope measurements were made in static mode, and each run consisted of 100 cycles with 4-s integration times, or 50 cycles at 8 s. Masses 142–150 were collected in cups L1 through H4, with mass 146 in the centre cup. Prior to each analysis, all masses were measured for ten 8-s cycles for blank subtractions. Blank corrections were negligible owing to effective washout of the previous sample. An Nd isotopic standard (for example, Ames Nd, JNdi-1) was run after every fourth sample as a consistency standard and to monitor accuracy and precision. All Nd isotopes were measured while monitoring masses 147 and 149 (Sm) allowing for interference corrections on 144, 148 and 150 (Nd). These corrections are negligible because of very small ^{147}Sm and ^{149}Sm intensities (0.01% of signal). Instrumental mass discrimination was corrected relative to ^{146}Nd/^{144}Nd = 0.7219 using an exponential law. All data have been normalized to La Jolla = 0.511858 or JNdi-1 = 0.512115; the average ^{147}Sm/^{144}Nd ratio of samples from each site was used to make the age-dependent correction for the in-growth of radiogenic ^{143}Nd since the time of deposition (see Source Data for Fig. 2).

REE+Y analyses. Sub-samples containing single fossil fish teeth were separated from cleaned samples for rare earth element and yttirium (REE+Y) concentration measurements. Samples for concentration determinations were weighed on a microbalance and dissolved in a solution containing 2% HNO_3 spiked with 2 p.p.b. (parts per billion) indium (In). Samples were measured on an Element2 ICP-MS at the University of South Carolina. All samples were measured in high-resolution mode to avoid oxide interferences from polyatomic species containing barium and REEs. Concentrations of REE+Y were obtained by external calibration using a mixed-element solution and a calibrated phosphate reference material, fossil bone composite[34]. The blank was incorporated into the calibration curve by measuring the 2% HNO_3 solution in which the samples were dissolved as a calibration blank.

Oligocene polar front. The past position of the polar front, which separates cold polar waters from warmer waters, is based on microfossil assemblage and lithological data from DSDP cores in the New Zealand sector of the Southern Ocean[19]. This approach is based on core top assemblages showing a clear relationship between calcareous and siliceous microfossils, relative to the position of present-day frontal zones. In this approach, the reconstructed polar front for a given time slice is drawn to the south of where calcareous microfossils are abundant (indicating that surface waters were relatively warm), and to the north of where siliceous microfossils are abundant (indicating that surface waters were relatively cold). Nelson and Cooke[19] use data from Deep Sea Drilling Program (DSDP) sites 267 & 268, 269, 280 & 281, 274 & 278 to determine the Oligocene (30-Ma) polar front. This position of the Oligocene polar front (Fig. 1) was incorporated into our tectonic reconstruction by back-tracking these sediment cores to their 30-Ma locations and redrawing the front. The specific information used to restore the Oligocene polar front is tabulated in Extended Data Table 1.

Endmember water-mass ε_{Nd} compilations and local sediment sources. The Palaeogene distribution of seawater Nd isotopes in the Pacific and the Atlantic–Indian sectors of the Southern Ocean has been reconstructed at medium to high resolution using fossil fish teeth from deep-sea sediment cores. The ε_{Nd} value of Pacific seawater has not demonstrated significant variability during the time interval investigated in this study[35,36]. Owing to the low resolution of the ferromanganese (Fe–Mn) Nd isotope records, we have chosen to use a newly developed fossil fish tooth Nd isotope stack from the equatorial Pacific—IODP (Integrated Ocean Drilling Program) sites U1331, U1332, U1333, U1334 and U1335[37]—as the Pacific endmember. The equatorial Pacific sites ranged from 2,500 to 4,500 m depth and 8° N to 2° S during the late Eocene to late Oligocene. The ε_{Nd} value for the Atlantic–Indian sectors of the Southern Ocean are compiled from ODP sites 689, 738, 757 and 1090 (refs 33, 38–42). The Nd isotope records used to draw the endmember fields in Fig. 2 are displayed in Extended Data Fig. 2. Information about the drill cores and Nd isotope data sources used to generate the endmember ε_{Nd} envelopes in Fig. 2, as well as the data sources for the local terrigenous inputs, are listed in Extended Data Table 3. The range of ε_{Nd} values carried by dissolved and particulate phases in the Murray River basin[43,44] is drawn on the right side of the figures.

Bulk-sediment geochemistry. To aid our interpretation of Pacific–Indian communication, we used a bulk-sediment aluminium/titanium (Al/Ti) record from the East Tasman Plateau to distinguish between mafic (that is, Pacific) sediment sources (Al/Ti < 9) and Indian sources, with Al/Ti ratios closer to that of average continental crust (Al/Ti = 15.8)[45]. A prominent shift from mafic sources to upper crustal values accompanies the Nd isotope shift and the timing of the northward migration of the South Tasman Rise margin north of the Oligocene polar front.

Age models. Age models for ODP sites 1124, 1168 and 1172 are from refs 46–48.

Subsidence models. The subsidence curves shown in Fig. 2a are based upon microfossil and facies analysis of the ODP Leg 189 sediment cores[13]. As noted in ref. 13, there is a significant subsidence anomaly for the East Tasman Plateau, resulting from conflicting palaeodepth estimates from site 1172 and the Cascade Guyot[13,49–51] (also called Cascade seamount in the literature). As shown in Extended Data Fig. 4, this discrepancy is ~1,900 m.

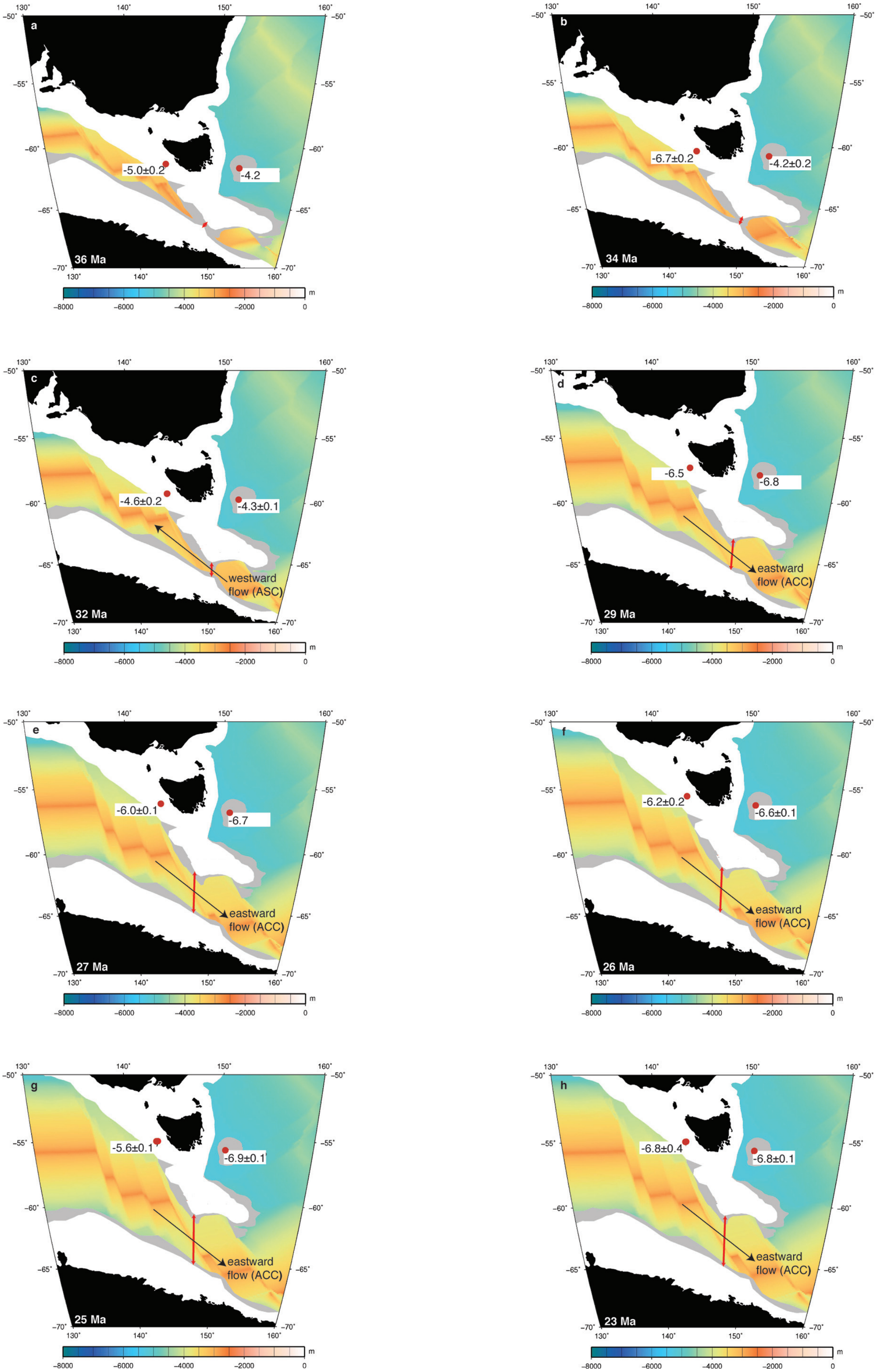
An alternative interpretation is that shallow marine sediments below 356 m in site 1172 were first deposited on the top/flank of the Cascade Guyot and subsequently transported to site 1172. We note that a 5-m section in ODP Site 1172 (361 to 356 metres below sea level), interpreted as shallow marine facies by Stickley *et al.*[11], has an up-section grain-size profile that is consistent with a classic Bouma sequence[52], and may actually represent turbidite sequences shed from the seamount when it was still above wave base and/or sea level. This possibility was discussed in ref. 13.

Given the interpretation that shallow marine sediments were transported by turbidites, we constructed the alternative subsidence model for ODP site 1172 (shown in Fig. 2a) by adjusting the published curve from ref. 13. We deepen the published subsidence curve between 38 Ma and 34 Ma by 1,900 m to account for the transport of sediments from the Cascade Guyot to site 1172. There is a sudden change in facies in site 1172 at 357 m to fully offshore pelagic sedimentation, dated to 30.2 Ma (ref. 11). We interpret this transition as the cessation of turbidite transport to site 1172, possibly due to the subsidence of the Cascade Seamount beneath sea level and/or wave base. We use the published subsidence curve from ref. 13 for times younger than 30 Ma. For the poorly constrained period between 34 Ma and 30 Ma, we linearly interpolate the subsidence.

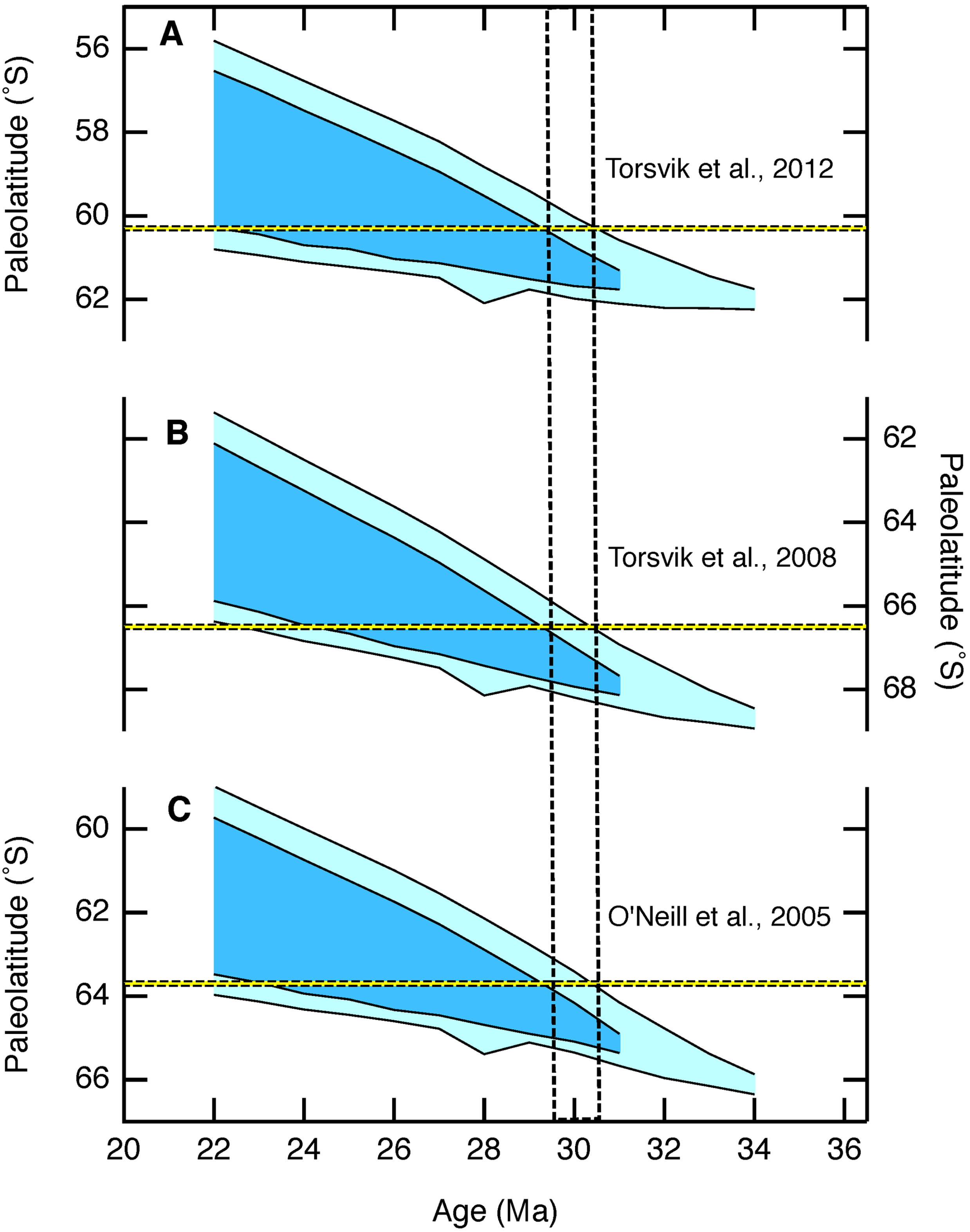
Our revised subsidence curve reconciles the available data from ODP site 1172 and the Cascade Guyot. The East Tasman Plateau is located to the east of Tasmania, a passive margin formed when continental break-up occurred at ~83 Ma between Australia and the Lord Howe Rise. Our revised subsidence curve is more consistent with the location of the East Tasman Plateau in such a tectonic setting, where subsidence rates some 50 million years after rifting are predicted to be low. Emplacement of the seamount yields thepotential for subsidence anomalies in the late Eocene related to volcanism and subsequent crustal loading from the mass of the seamount.

31. Boyle, E. A. Cadmium, zinc, copper and barium in foraminifera tests. *Earth Planet. Sci. Lett.* **53,** 11–35 (1981).
32. Boyle, E. A. & Keigwin, L. D. Comparison of Atlantic and Pacific paleochemical records for the last 215,000 years; changes in deep ocean circulation and chemical inventories. *Earth Planet. Sci. Lett.* 76, 135–150 (1985/86).
33. Scher, H. D. & Delaney, M. L. Breaking the glass ceiling for high resolution Nd isotope records in early Cenozoic paleoceanography. *Chem. Geol.* **269,** 329–338 (2010).
34. Chavagnac, V. *et al.* Towards the development of a fossil bone geochemical standard: an inter-laboratory study. *Anal. Chim. Acta* **599,** 177–190 (2007).
35. Ling, H. F. *et al.* Evolution of Nd and Pb isotopes in Central Pacific seawater from ferromanganese crusts. *Earth Planet. Sci. Lett.* **146,** 1–12 (1997).
36. Ling, H.-F. *et al.* Differing controls over the Cenozoic Pb and Nd isotope evolution of deepwater in the central North Pacific Ocean. *Earth Planet. Sci. Lett.* **232,** 345–361 (2005).
37. Scher, H. Stacking PEAT; a neodymium isotope stack for the Paleogene equatorial Pacific. *Rendiconti Soc. Geol. Ital.* **31,** 191–192 (2014).
38. Scher, H. D. & Martin, E. E. Circulation in the Southern Ocean during the Paleogene inferred from neodymium isotopes. *Earth Planet. Sci. Lett.* **228,** 391–405 (2004).
39. Scher, H. D., Bohaty, S., Zachos, J. C. & Delaney, M. L. Two-stepping into the icehouse: East Antarctic weathering during progressive ice-sheet expansion at the Eocene-Oligocene Transition. *Geology* **39,** 383–386 (2011).

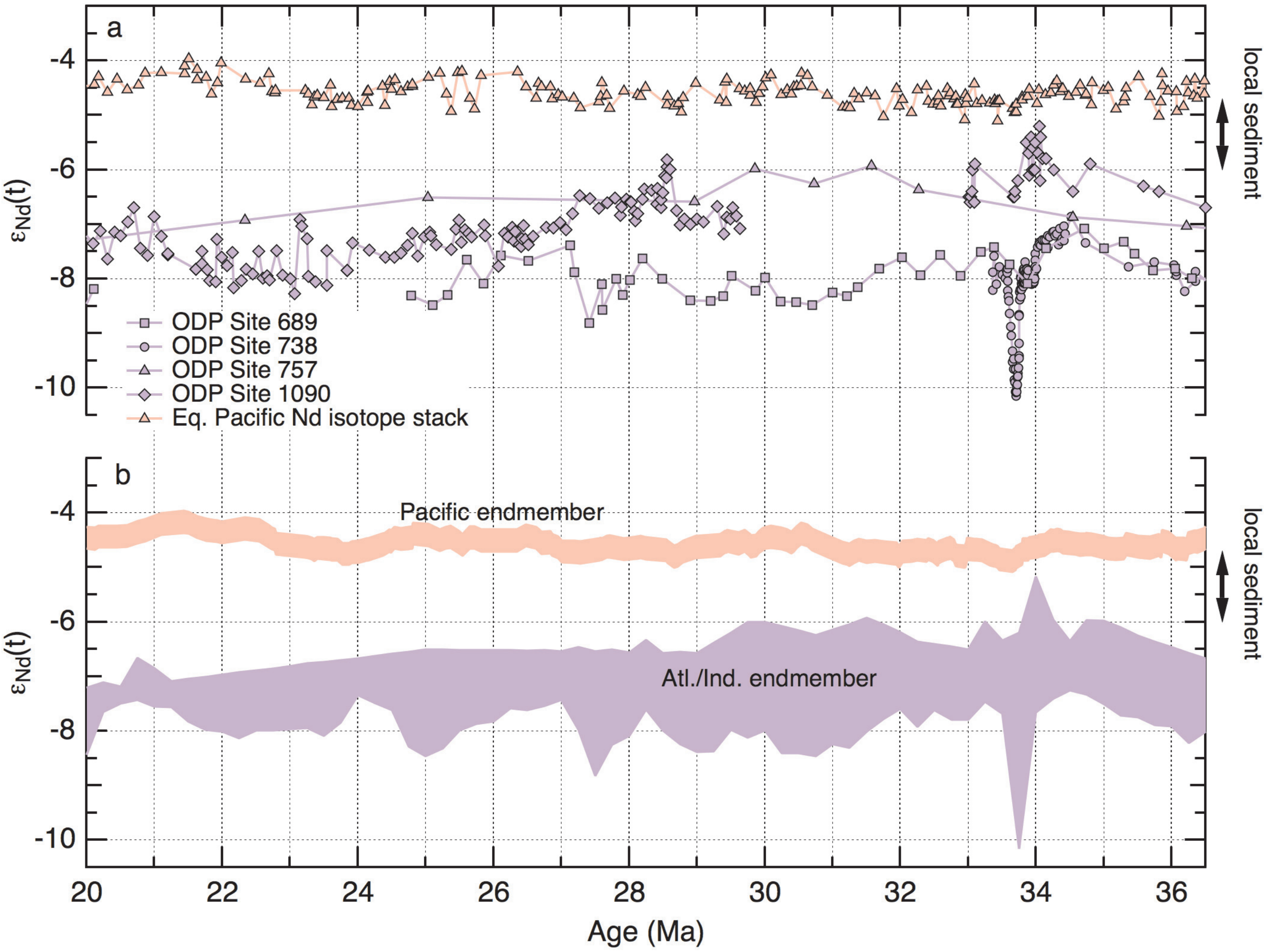
40. Scher, H. D., Bohaty, S. M., Smith, B. & Munn, G. Isotopic interogation of a suspected late Eocene glaciation. *Paleoceanography* **29,** 628–644 (2014).
41. Martin, E. E. & Scher, H. A Nd isotopic study of southern sourced waters and Indonesian throughflow at intermediate depths in the Cenozoic Indian Ocean. *Geochem. Geophys. Geosyst.* **7,** Q09N02 (2006).
42. Scher, H. D. & Martin, E. E. Timing and climatic consequences of the opening of Drake Passage. *Science* **312,** 428–430 (2006).
43. Goldstein, S. J. & Jacobsen, S. B. The Nd and Sr isotopic systematics of river-water dissolved material — implications for the sources of Sd and Sr in seawater. *Chem. Geol.* **66,** 245–272 (1987).
44. Goldstein, S. J. & Jacobsen, S. B. Nd and Sr isotopic systematics of river water suspended material; implications for crustal evolution. *Earth Planet. Sci. Lett.* **87,** 249–265 (1988).
45. Taylor, S. R. & McLennan, S. M. *The Continental Crust: Its Composition and Evolution* (Blackwell, 1985).
46. Joseph, L. H., Rea, D. K. & van der Pluijm, B. A. Neogene history of the Deep Western Boundary Current at Rekohu sediment drift, Southwest Pacific (ODP Site 1124). *Mar. Geol.* **205,** 185–206 (2004).
47. Stickley, C. E. *et al.* Late Cretaceous to Quaternary biomagnetostratigraphy of ODP Sites 1168, 1170, 1171, and 1172, Tasmanian Gateway. *Proc. ODP Sci. Res.* http://www-odp.tamu.edu/publications/189_SR/111/111.htm (2004).
48. Fuller, M. & Touchard, Y. in *The Cenozoic Southern Ocean. Tectonics, Sedimentation, and Climate Change between Australia and Antarctica* (eds Exon, N., Kennett, J. P. & Malone, M.) 63–78 (American Geophysical Union, 2004).
49. Lanyon, R., Varne, R. & Crawford, A. J. Tasmanian tertiary basalts, the Balleny plume, and opening of the Tasman sea (southwest Pacific-ocean). *Geology* **21,** 555–558 (1993).
50. Quilty, P. G. Late Eocene foraminifers and palaeoenvironment, Cascade Seamount, southwest Pacific Ocean: implications for seamount subsidence and Australia-Antarctica Eocene correlation. *Austral. J. Earth Sci.* **48,** 633–641 (2001).
51. Quilty, P. G. Eocene and younger biostratigraphy and lithofacies of the Cascade Seamount, East Tasman Plateau, southwest Pacific Ocean. *Austral. J. Earth Sci.* **44,** 655–665 (1997).
52. Bouma, A. H. *Sedimentology of Some Flysch Deposits: A Graphic Approach to Facies Interpretation* (Elsevier, 1962).
53. Müller, R. D., Sdrolias, M., Gaina, C. & Roest, W. R. Age, spreading rates, and spreading asymmetry of the world's ocean crust. *Geochem. Geophys. Geosyst.* **9,** Q04006 (2008).
54. Torsvik, T. H. *et al.* Phanerozoic polar wander, palaeogeography and dynamics. *Earth Sci. Rev.* **114,** 325–368 (2012).
55. Torsvik, T. H., Steinberger, B., Cocks, L. R. M. & Burke, K. Longitude: linking Earth's ancient surface to its deep interior. *Earth Planet. Sci. Lett.* **276,** 273–282 (2008).



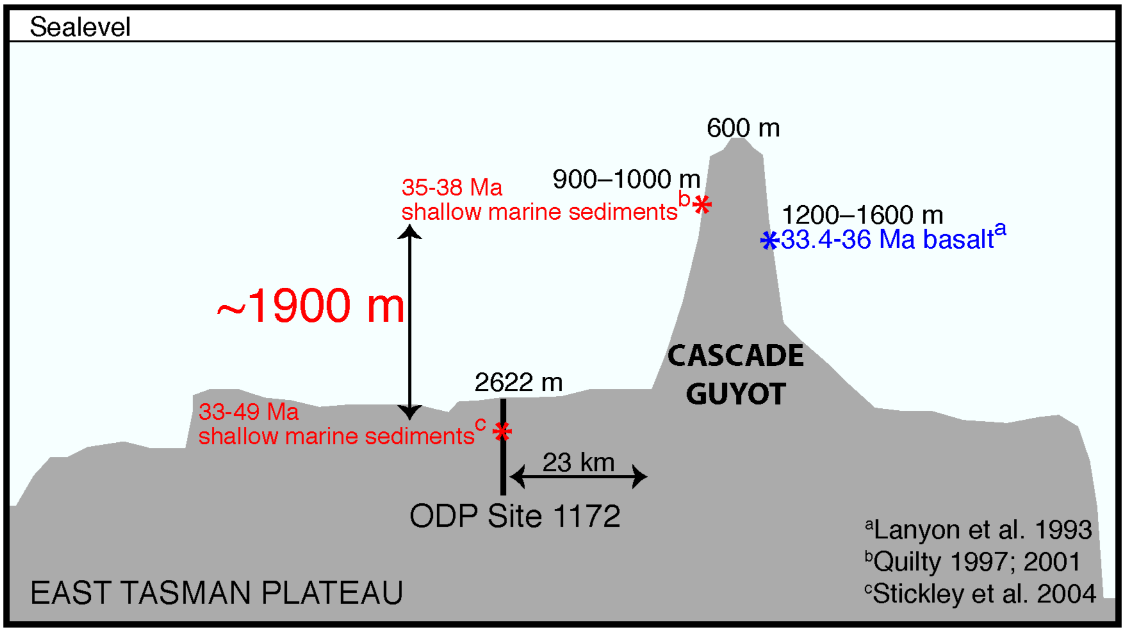
Extended Data Figure 1 | Tectonic reconstructions of the Tasmanian Gateway. Present-day coastlines (black), continental shelves (white), and range of the innermost and outermost continent–ocean boundaries (COBs) for the conjugate margins of the Tasmanian Gateway (grey) are shown for time slices from the late Eocene (**a**, **b**) to the early Oligocene (**c**–**h**). The palaeobathymetry grid is based on ref. 53. Palaeolatitudes of the innermost and outermost COB locations in Fig. 2 are based on their position along a meridional section drawn through the narrowest portion of the gateway (red arrow in each reconstruction). Palaeocurrent direction (indicated by black arrows) is based on the Nd isotope data from sites 1168 and 1172 (red circles), which indicate westward or eastward flow, depending on the ε_{Nd} values at the sites. The ε_{Nd} values listed on the diagrams are the average values for the time slices shown. Numbers in parentheses show the standard error (see Source Data).



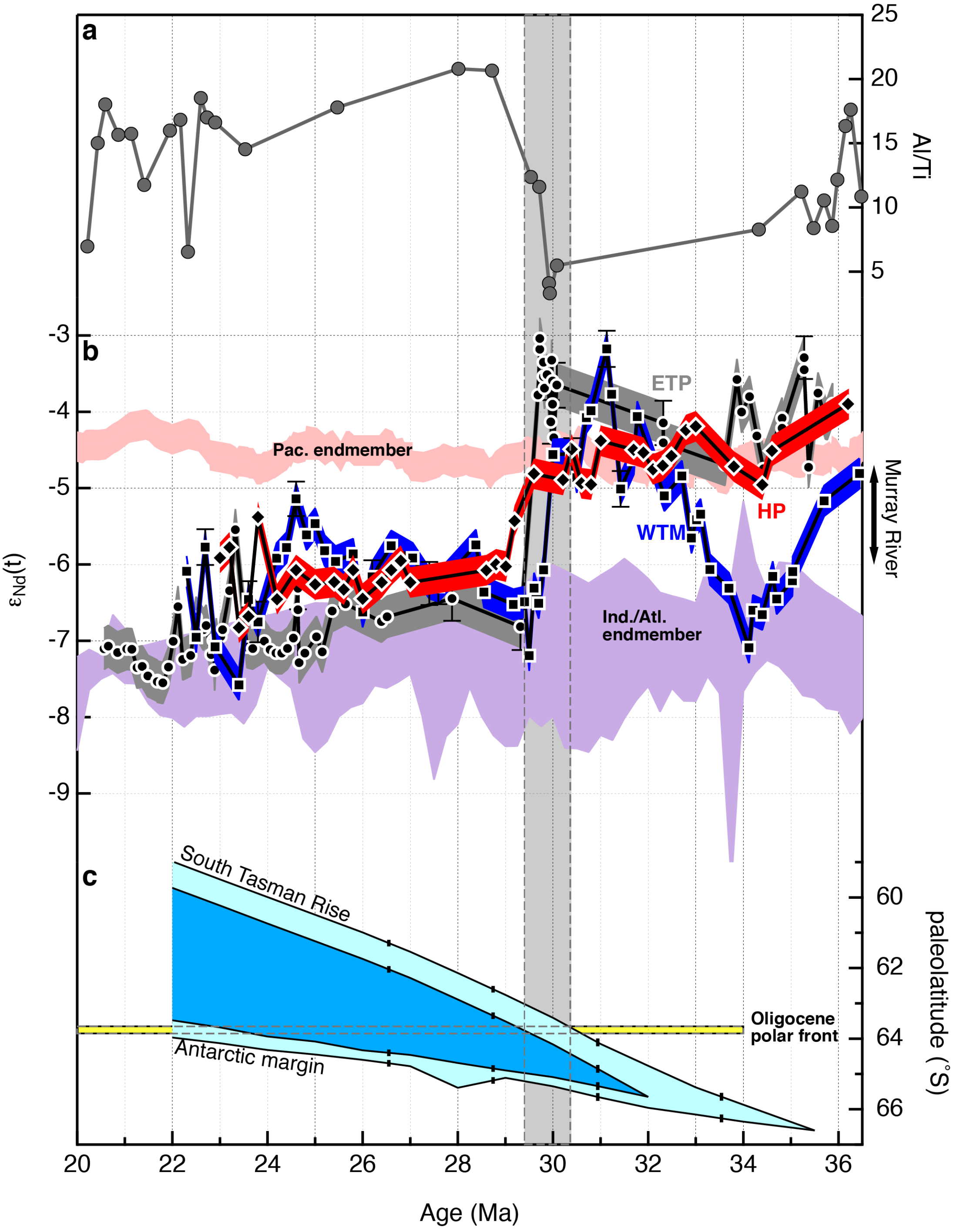
Extended Data Figure 2 | Comparison of absolute reference frames. The relative position of the northern conjugate margin of the Tasmanian Gateway relative to the Oligocene polar front was tested using three reference frames: **a**, ref. 54; **b**, ref. 55; and **c**, ref. 18. Although the absolute palaeolatitudes of these features differ between these reference frames, the timing of the northern margin transit across the Oligocene polar front is not affected by the choice of reference frame because the reconstruction of the Oligocene polar front is reconstructed using the same rotation data as the margins. In each reference frame, the northern margin of the gateway migrates across the Oligocene polar front between 29 and 30 Ma. Light blue bands correspond to the maximum gateway width based on geophysically determined positions of the innermost COBs on the South Tasman Rise and Antarctica[8], selected along the narrowest meridional transect of the gateway (Extended Data Fig. 1). Dark blue bands show the minimum gateway widths based on the outermost geophysically determined COBs. The Oligocene position of the polar front (yellow band) was derived from microfossil assemblage data in deep-sea sediment cores[19]; these data were incorporated into the plate reconstruction used in this study. The grey band centred at 30 Ma highlights the age range of the South Tasman Rise crossing the Oligocene polar front.



Extended Data Figure 3 | Water-mass endmembers used for this study. Nd isotope records used to demarcate Pacific (red) and Indian–Atlantic (purple) waters over the study interval. **a**, Individual Nd isotope records (see Extended Data Table 3 for reference list). **b**, Range of the endmember waters for the Pacific and Indian–Atlantic. Envelope around the Pacific Nd isotope stack is instrumental uncertainty (2σ).



Extended Data Figure 4 | Anomalous subsidence histories of Cascade Guyot and East Tasman Plateau. Cascade Guyot subsided at least 1,000 m since the late Eocene, judging by the sediments dredged high on the flank. Shallow marine sediments are currently ~2,900 m below sea level at ODP site 1172, suggesting that the East Tasman Plateau subsided three times more than the Guyot over the same interval.



Extended Data Figure 5 | Bulk sediment geochemistry. **a**, Detrital Al/Ti ratios from ODP site 1172 (red triangles). **b**, Fossil fish tooth Nd isotope records from the Western Tasmanian Margin (WTM; site 1168; blue), the East Tasman Plateau (ETP; site 1172; grey), and the Hikurangi Plateau (HP; site 1124; red). Error envelopes represent instrumental uncertainty (2σ) based on replicate Nd isotope analyses of the JNdi-1 Nd isotope standard. Error bars (standard errors) are shown for samples where the standard error of the measurement exceeds 2σ instrumental uncertainty. The Pacific and Indian–Atlantic ε_{Nd} envelopes are based on fossil fish tooth Nd isotope records from the equatorial Pacific, southern Kerguelen Plateau, Ninetyeast ridge, and Maud Rise (Extended Data Fig. 3). The range of ε_{Nd} values of weathering products in the Murray River is shown on the right axis. **c**, The palaeolatitudes of the conjugate margins of the Tasmanian Gateway outline the width and position of the gateway during its progressive opening. Light blue bands correspond to the maximum gateway width based on geophysically determined positions of the innermost COBs on the South Tasman Rise and Antarctica[8], selected along the narrowest meridional transect of the gateway (Extended Data Fig. 1). The dark blue band shows the minimum gateway widths based on the outermost geophysically determined COBs. Error bars (standard errors) on the northern and southern boundaries of the gateway reflect uncertainty in the positions of the conjugate margins based on the 95% confidence intervals from the rotations[17]. The Oligocene position of the polar front (yellow band) was derived from microfossil assemblage data in deep-sea sediment cores[19], which were incorporated into the plate reconstruction used in this study. The grey band centred at 30 Ma highlights the age range of the South Tasman Rise crossing the Oligocene polar front.

Extended Data Table 1 | Lithological information from DSDP cores used to restore the Oligocene polar front, based on data in ref. 19

DSDP Site	lithology (major, minor)	relative surface water temperature interpretation
267	carbonate	warm
268	terrigenous, minor opal	cold
269	terrigenous	cool
274	terrigenous and opal	cold
278	opal, minor carbonate	cold
280	opal, minor terrigenous and carbonate	cold
281	terrigenous/carbonate	cool

Extended Data Table 2 | Ocean Drilling Program site information for this study

Site name; province, Ocean Basin	Latitude, Longitude	Present depth (m)	Age model applied
ODP Site 1124; Rehoku Drift, Pacific	39°29.897'S, 176°31.893'W	3966.8	Joseph et al., 2004
ODP Site 1168; west Tasman margin, Indian	42°36.568'S, 144°24.761'E	2463.6	Stickley et al., 2004
ODP Site 1172; east Tasman Plateau, Pacific	43°57.571'S, 149°55.706'E	2621.9	Fuller and Touchard, 2004

References are Joseph *et al.*[46], Stickley *et al.*[47] and Fuller and Touchard[48].

Extended Data Table 3 | Details of the ODP sites used as the Pacific and Indian–Atlantic endmembers in Fig. 1 and Extended Data Table 2

Site/River	Location	Latitude	Longitude	Paleodepth (m)[a]	Nd isotope data source
ODP Site 689	Maud Rise, Atlantic sector Southern Ocean	64.5˚S	5.1˚W	1500	Scher and Martin, 2004
ODP Site 738	Southern Kerguelen Plateau, Indian sector Southern Ocean	62.7˚S	82.8˚E	1750	Scher et al., 2011
ODP Site 757	Ninetyeast Ridge, Indian Ocean	17˚S	88.2˚E	1000-1500	Martin and Scher, 2006
ODP Site 1090	Aghulas Ridge, south Atlantic Ocean	42.9˚S	8.9˚E	3400	Scher and Martin, 2006; 2008
IODP sites U1331, U1332, U1333, U1334, U1335	Equatorial Pacific	5 - 12˚N	126 - 142˚W	2500-4500	Scher, 2014
Murray River	Murray Bridge, SA	na	na	na	Goldstein and Jacobsen, 1987; Goldstein and Jacobsen, 1988

[a] Palaeodepth at 34 Ma.

doi:10.1038/nature14472

A Middle Triassic stem-turtle and the evolution of the turtle body plan

Rainer R. Schoch[1] & Hans-Dieter Sues[2]

The origin and early evolution of turtles have long been major contentious issues in vertebrate zoology[1–11]. This is due to conflicting character evidence from molecules and morphology and a lack of transitional fossils from the critical time interval. The ~220-million-year-old stem-turtle *Odontochelys* from China[12] has a partly formed shell and many turtle-like features in its postcranial skeleton. Unlike the 214-million-year-old *Proganochelys* from Germany and Thailand, it retains marginal teeth and lacks a carapace. *Odontochelys* is separated by a large temporal gap from the ~260-million-year-old *Eunotosaurus* from South Africa, which has been hypothesized as the earliest stem-turtle[4,5]. Here we report a new reptile, *Pappochelys*, that is structurally and chronologically intermediate between *Eunotosaurus* and *Odontochelys* and dates from the Middle Triassic period (~240 million years ago). The three taxa share anteroposteriorly broad trunk ribs that are T-shaped in cross-section and bear sculpturing, elongate dorsal vertebrae, and modified limb girdles. *Pappochelys* closely resembles *Odontochelys* in various features of the limb girdles. Unlike *Odontochelys*, it has a cuirass of robust paired gastralia in place of a plastron. *Pappochelys* provides new evidence that the plastron partly formed through serial fusion of gastralia[3,13]. Its skull has small upper and ventrally open lower temporal fenestrae, supporting the hypothesis of diapsid affinities of turtles[2,7–10,14,15].

Turtles are readily diagnosed by the possession of a bony shell consisting of a dorsal carapace and a ventral plastron, which are linked by bony bridges on either side. In recent years embryological studies on extant turtles[1,3] and recognition of stem-turtles[4–6,12] have provided evidence about the stepwise acquisition of features of this unique body plan. The discovery of a new stem-turtle from the Middle Triassic (Ladinian) of Germany sheds new light on this evolutionary transition as well as on the long-contentious relationships of turtles to other amniotes.

Reptilia Laurenti, 1768
Pan-Testudines Joyce, Parham and Gauthier, 2004 (ref. 16)
Pappochelys gen. nov.

Etymology. *Pappos* (Greek): grandfather; *chelys* (Greek): turtle. Type species. *Pappochelys rosinae*.

Pappochelys rosinae sp. nov.

Etymology. In honour of I. Rosin, who prepared key specimens of the new taxon.
Holotype. Staatliches Museum für Naturkunde Stuttgart, SMNS 91360, incomplete, partly articulated postcranial skeleton (Fig. 1a, b).
Referred material. SMNS 90013, disarticulated skeleton with incomplete skull (Fig. 1c, d), and 18 additional specimens. See Supplementary Information for details.
Type locality. Schumann quarry, Eschenau, Vellberg municipality, Baden-Württemberg, Germany.
Type horizon. Top of Untere Graue Mergel, Lower Keuper (Erfurt Formation); late Middle Triassic (Ladinian: Longobardian). The fossils occur in a 5- to 15-cm-thick layer of dark grey lacustrine claystone, along with fishes, temnospondyl stem-amphibians, and mostly terrestrial diapsid reptiles[17]. Despite extensive collecting in Lower Keuper strata for the past two centuries, not a single diagnostic bone of the new reptile was discovered until recently. Since 2006, the aforementioned fossils referable to this taxon have been recovered from the type locality.
Diagnosis. Small (estimated adult length ~20 cm); skull proportionately small, with short, deep temporal region; parietal with distinct occipital flange; jaw bones bearing teeth; squamosal and parietal bounding much of upper temporal fenestra; lower temporal opening

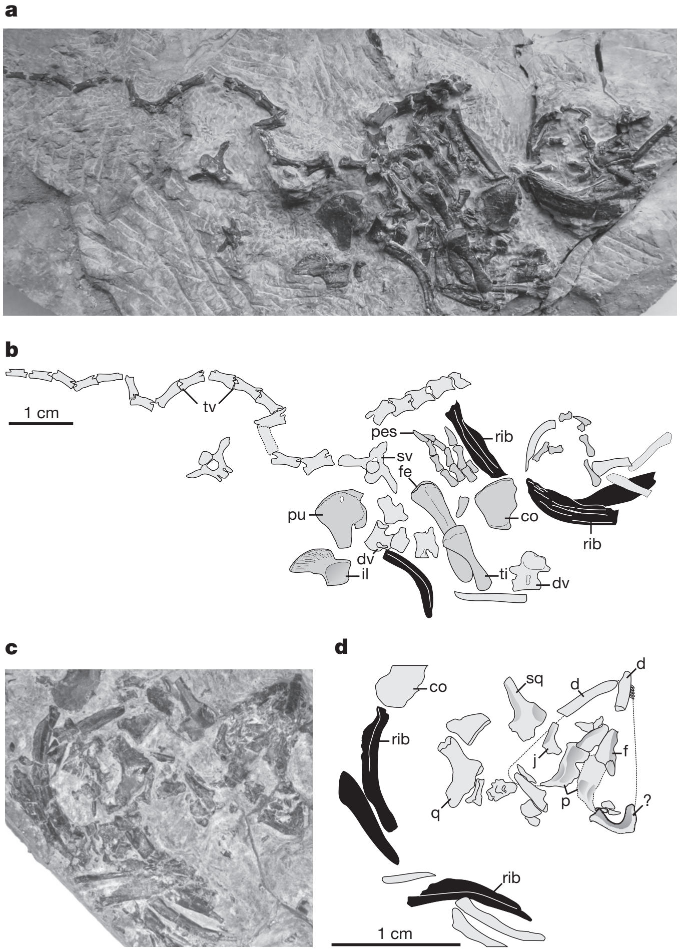


Figure 1 | *P. rosinae*. **a**, **b**, Articulated partial postcranial skeleton (SMNS 91360, holotype); **c**, **d**, disarticulated skeleton with incomplete skull (SMNS 90013). Photographs with explanatory outline drawings. Trunk ribs highlighted in black. Abbreviations: co, coracoid; d, dentary; dv, dorsal vertebra; f, frontal; fe, femur; il, ilium; j, jugal; p, parietal; pu, pubis; q, quadrate; sq, squamosal; sv, sacral vertebra; ti, tibia; tv, tail vertebra; question mark, unidentified cranial bone.

[1]Staatliches Museum fur Naturkunde Stuttgart, Rosenstein 1, D-70191 Stuttgart, Germany. [2]Department of Paleobiology, National Museum of Natural History, MRC 121, PO Box 37012, Washington, District of Columbia 20013-7012, USA.

open ventrally; trunk ribs anteroposteriorly broad, T-shaped in cross-section and with dorsal sculpturing; gastralia paired, robust, with ridged external surface; whip-like tail comprising some 50% of total length; scapula with tall, straight dorsal process and small 'acromial' flange; coracoid plate-like; ilium with long postacetabular process; pubis with thyroid foramen and distinct lateral process.

All specimens of the new taxon were mechanically prepared. Histological work on fossil bones employed standard techniques used for making petrographic thin-sections.

The skull of *Pappochelys* (Fig. 2c) has a large orbit and short, pointed snout and is triangular in dorsal view. Its broad cheek region has a small rounded upper and a ventrally open lower temporal fenestra. The configuration of the bones in the temporal region is consistent with that in diapsid reptiles.

The premaxilla, maxilla, and dentary bear teeth. The teeth are peg-like in the maxilla (Fig. 2a, b) and slightly inclined posteriorly in the anterior part of the dentary (Fig. 2i). In both upper and lower jaws, the more posterior teeth are smaller and somewhat more robust. The premaxilla bears four teeth, the maxilla up to 17, and the dentary at least 29. Tooth implantation appears to be subthecodont. The short maxilla has a large, posterodorsally directed facial process and a short anterior ramus. Its posterior process decreases in height posteriorly. The frontal (Extended Data Fig. 1g) is longer than the nasal and parietal. The parietal (Fig. 2d, Extended Data Fig. 1c, d) has a short posterolateral wing and distinctly offset occipital flange. There is no trace of a pineal foramen. The postorbital (Fig. 2e and Extended Data Fig. 1e) is short anteroposteriorly, with a rounded posterior and a long ventral process contacting the posterior margin of the dorsal process of the jugal. The postfrontal (Extended Data Fig. 1f) is triangular. The slender jugal (Fig. 2h and Extended Data Fig. 2) has an elongate, anterodorsally curving anterior process, a tapering dorsal process, and a short posterior process that indicates a ventrally largely open lower temporal fenestra. The squamosal (Fig. 2f) has a long ventral process, tapering postorbital ramus, rounded posterior buttress, and median parietal process. The massive quadrate (Fig. 2g) has a slightly concave posterior margin. The palate is not clearly exposed in any available specimen. The dentary (Fig. 2i and Extended Data Fig. 2) is slender and curves anterodorsally. No other mandibular bones have been identified so far.

Cervical vertebrae (Fig. 3g) are elongate and low, with extensive postzygapophyses and low neural spines. The number of presacral vertebrae and relative lengths of the neck and trunk remain unknown. Dorsal vertebrae (Fig. 3h, i) have long cylindrical centra with nearly vertical rib facets at about midlength. The neural arch is fused to the centrum in adults, and the neural spine is low. There is no trace of neurals. Although the trunk region is disarticulated in all available specimens, the maximum number of trunk vertebrae did not exceed nine. The anteroposteriorly broad trunk ribs (Fig. 3a–d) each bear slightly asymmetrical anterior and posterior flanges and are distinctly T-shaped in cross-section. The confluent rib heads have a figure-eight shape in end view. The distal ends of the ribs are tapered. Undistorted

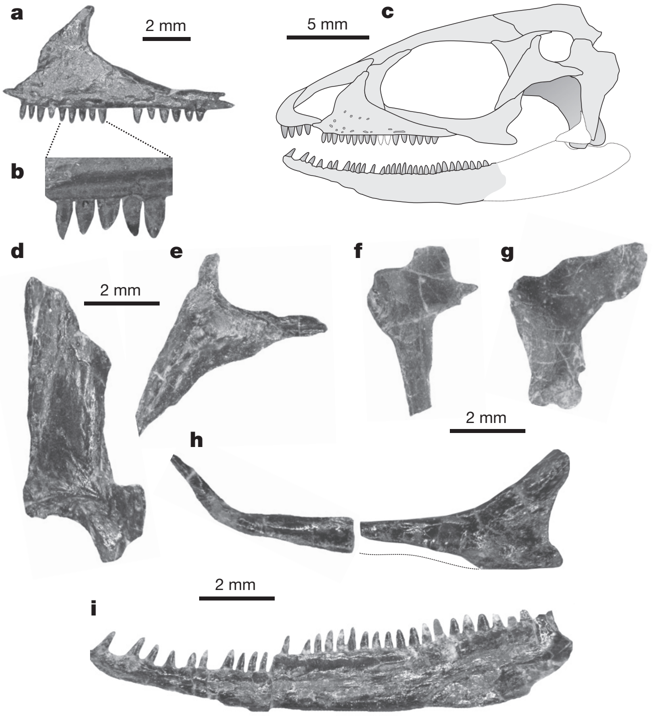


Figure 2 | Skull elements of *P. rosinae* (digitally extracted from surrounding matrix). **a**, **b**, Left maxilla (SMNS 91431; **a**, labial view; **b**, lingual view of marked section); **c**, skull reconstruction in lateral view, with preserved elements indicated in grey; **d**, right parietal (SMNS 91356); **e**, right postorbital (SMNS 91356); **f**, right squamosal (SMNS 90013); **g**, right quadrate (SMNS 90013); **h**, left jugal (SMNS 92066, broken into two segments and partly preserved as an impression); **i**, left dentary (SMNS 92066).

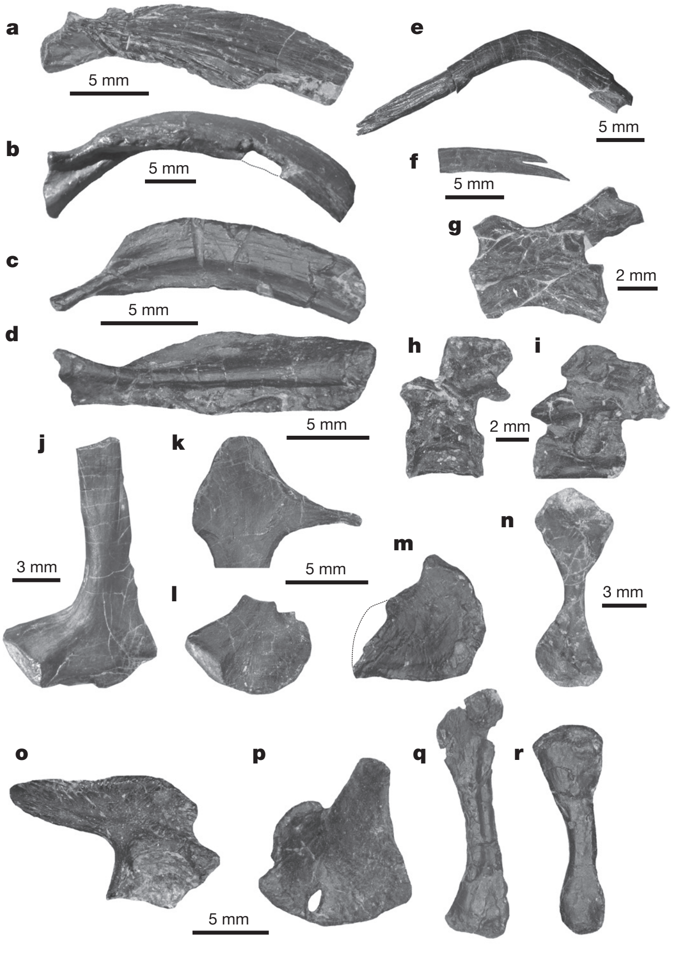


Figure 3 | Postcranial elements of *P. rosinae* (digitally extracted from surrounding matrix). **a**, Trunk rib in dorsal view (SMNS 92067); **b**, trunk rib in anterior view (SMNS 92068); **c**, **d**, trunk ribs in ventral view (**c**, SMNS 92063; **d**, SMNS 91360); **e**, gastralium (SMNS 91360); **f**, lateral ends of two fused gastralia (SMNS 91363); **g**, cervical vertebra (SMNS 91360); **h**, anterior dorsal vertebra (SMNS 91356); **i**, mid-dorsal vertebra (SMNS 91360); **j**, right scapula (SMNS 92044); **k**, incomplete interclavicle (SMNS 91895); **l**, proximal portion of right scapula (SMNS 91895); **m**, right coracoid (SMNS 91360); **n**, left humerus (SMNS 91113); **o**, right ilium (SMNS 91895); **p**, left pubis (SMNS 91360); **q**, right femur (SMNS 91356); **r**, left tibia (SMNS 91360).

ribs curve laterally, with the posterior flange being more extensive than the anterior one. Large ribs bear pronounced sculpturing on the dorsal surface, comprising ridges that may bear tubercles in places (Fig. 3a). This sculpturing suggests an intradermal origin of these structures[5]. Thin-sections of trunk ribs (Extended Data Fig. 4) show that they are rather compact and closely resemble those of *Eunotosaurus* in their histological structure[5]. The whip-like tail of *Pappochelys* comprises over 23 vertebrae with low neural arches and long, cylindrical centra (Fig. 1a, b).

The trunk has large paired gastralia (Fig. 3e and Extended Data Fig. 3). Individual elements are thick, with ridged surfaces and tapered ends. Successive gastralia occasionally fused to each other, as suggested by several particularly robust elements with forked distal ends (Fig. 3f). In ventral view, the anterior gastralia extend anterolaterally, whereas the reverse obtains on the posterior gastralia. None of the available fossils preserves undisturbed pairs of gastralia.

The scapula (Fig. 3j, l) has a tall, slender dorsal process, which was probably aligned vertically, and a short 'acromial' process. Its glenoid facet faces posterolaterally. The plate-like coracoid (Fig. 3m) has an anterolateral glenoid facet. The interclavicle (Fig. 3k) has a rounded anterior process, posterolaterally extending lateral processes contacting the clavicles, and a long, tapering posterior process. The robust humerus (Fig. 3n) has expanded articular ends. Radius and ulna are slender, and the manus has slender digits with long, narrow unguals.

The pelvis closely resembles those of *Odontochelys*[12] and *Proganochelys*[18]. However, the ischium remains separate from the pubis. The ilium (Fig. 3o) has a long postacetabular process with a strongly striated lateral surface and a straight dorsal margin. The pubis (Fig. 3p) has an oval thyroid foramen and a distinct lateral process. There is no trace of a hypoischium.

The hindlimb is only slightly longer than the forelimb. The S-shaped femur (Fig. 3q) has a distinct internal trochanter and an offset head. The robust tibia (Fig. 3r) is much shorter than the femur. The pes is slightly larger than the manus.

Morphologically *Pappochelys* represents an intermediate stage between *Eunotosaurus*[4,5,19] and *Odontochelys*[12]. It shares the T-shaped cross-sectional outline of the broad trunk ribs with both taxa. Saurosphargid reptiles also have broad trunk ribs[20,21], but these differ from those of stem-turtles in being flatter (with the shaft not offset), straighter, single-headed, and uniform throughout the trunk region. By contrast, the thoracic ribs of *Pappochelys* and *Odontochelys* vary in shape with respect to their position, and some ribs have asymmetrical anterior and posterior flanges. Unlike *Eunotosaurus* and *Pappochelys*, *Odontochelys* has neurals and supports the hypothesis that the turtle carapace developed by outgrowth of intramembranous bone from the periosteum of the ribs and neural spines. *Odontochelys* also has a fully developed plastron whereas *Eunotosaurus* has paired gastralia[4], as does *Pappochelys*. The gastralia, together with ventral elements of the shoulder girdle, formed the plastron in turtles[3,13]. Although *Pappochelys* lacks a plastron, the typically thickened and sometimes two-ended gastralia indicate increased ossification in the ventral region and incipient fusion of successive elements in this taxon, supporting the hypothesis that the turtle plastron partly formed through their co-ossification[3,13]. The lateral ends of the plastral elements in *Odontochelys* form spine-like projections[12] that resemble the distal ends of the gastralia in *Pappochelys* both in their alignment and in their striated surface texture.

The scapula and coracoid of *Pappochelys* closely resemble those of *Odontochelys* and the more derived *Proganochelys*. Compared with *Proganochelys*, the 'acromial' process of the scapula is short and forms a rounded anteromedial edge and the coracoid is not as expanded.

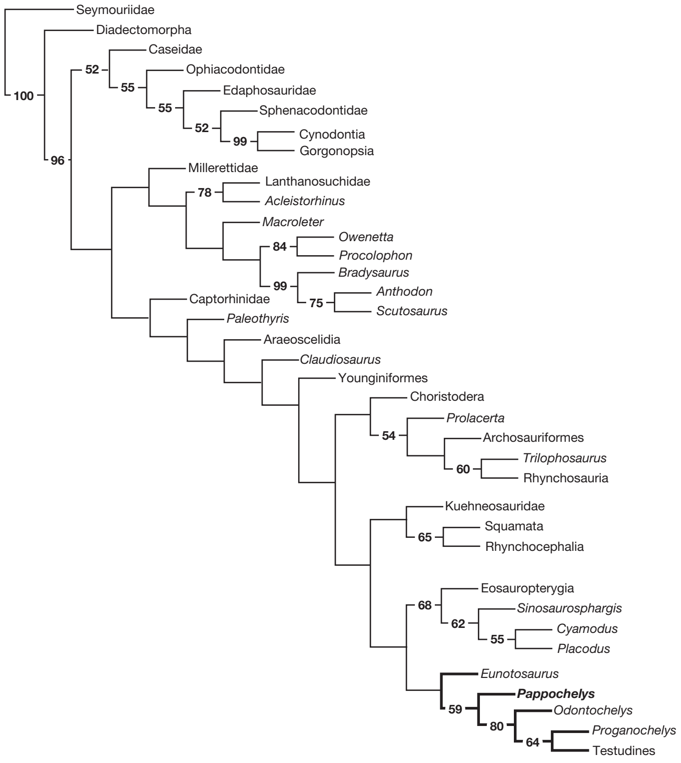


Figure 4 | Phylogenetic position of Pan-Testudines including *Pappochelys* among Amniota based on maximum parsimony analysis of the data matrix in ref. 5. Numbers at nodes indicate bootstrap percentages (only those >50) for each node. The tree has a length of 759 steps, a consistency index of 0.32, and a rescaled consistency index of 0.21.

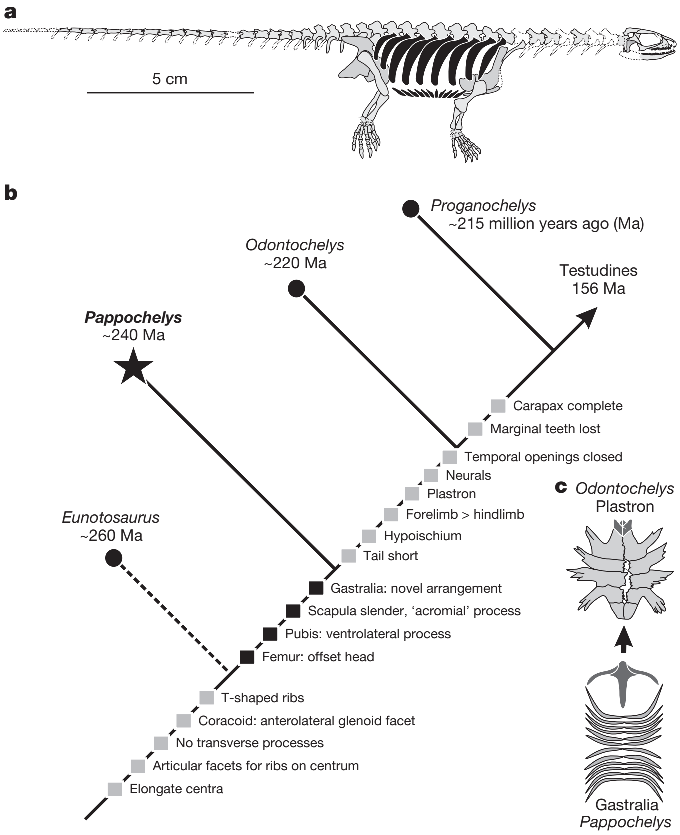


Figure 5 | Early evolution of the turtle body plan. **a**, Restoration of the skeleton of *Pappochelys* in lateral view (as yet unknown elements in white; preserved bones in grey; trunk ribs and gastralia highlighted in black); **b**, successive appearance of key features of the turtle body plan; **c**, plastron of *Odontochelys* and reconstructed ventral bones of the shoulder girdle and gastralia set in *Pappochelys* (elements of the shoulder girdle and their homologues are indicated in a darker shade of grey).

The pubis of *Pappochelys* has a distinct lateral process, as in *Odontochelys* and more derived turtles, where it contacts the plastron. The ilium closely resembles that of *Proganochelys* in outline.

The presence of two temporal openings on either side of the cranium in *Pappochelys* supports the hypothesis of diapsid affinities for turtles[2,14,15]. The configuration of the squamosal, postorbital, and parietal closely resembles the condition in diapsid reptiles, whereas the small size of the upper temporal fenestra may suggest incipient reduction of that opening. A phylogenetic position of turtles within Diapsida has been consistently recovered by phylogenetic analyses based on molecular data (although most of the latter specifically place turtles with or close to archosaurs)[7–10] but also by some morphologically based studies[2,14,15]. Our phylogenetic analysis (Fig. 4 and Extended Data Figs 5 and 6; for detailed data see Supplementary Information) recovered Pan-Testudines as the sister-taxon to Sauropterygia, a clade of marine saurian reptiles, and this grouping as the sister-group to Lepidosauriformes as previously suggested in refs 2 and 14. Traditionally, palaeontologists interpreted turtles as primarily lacking temporal openings and accordingly assigned them a basal position among reptiles[6]. Similarly, the skull of *Eunotosaurus* was long considered 'anapsid' but, like *Pappochelys*, it has ventrally open lower temporal openings[22]. Furthermore, new research indicates that *Eunotosaurus* has upper temporal openings concealed by large supratemporals[23].

Pappochelys is the most common reptile in the Vellberg lake deposit known so far, and is represented by various growth stages, which suggests that it either lived along the lakeshore or frequently entered the lake. Under a scenario that the turtle shell initially evolved in an aquatic setting[2], the plastron may have first developed as protection and 'bone ballast' for controlling buoyancy[6]. The thick gastralia and ribs in *Pappochelys* are consistent with aquatic or semi-aquatic habits. Although the oldest fully shelled turtles were probably terrestrial[24,25], *Odontochelys* apparently lived in deltaic or lagoonal settings along a coastline[6,12].

In summary, *Pappochelys* provides a new stage in the evolution of the turtle body plan (Fig. 5) and critical evidence for the diapsid relationships of turtles.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 8 January; accepted 13 April 2015.

Published online 24 June 2015.

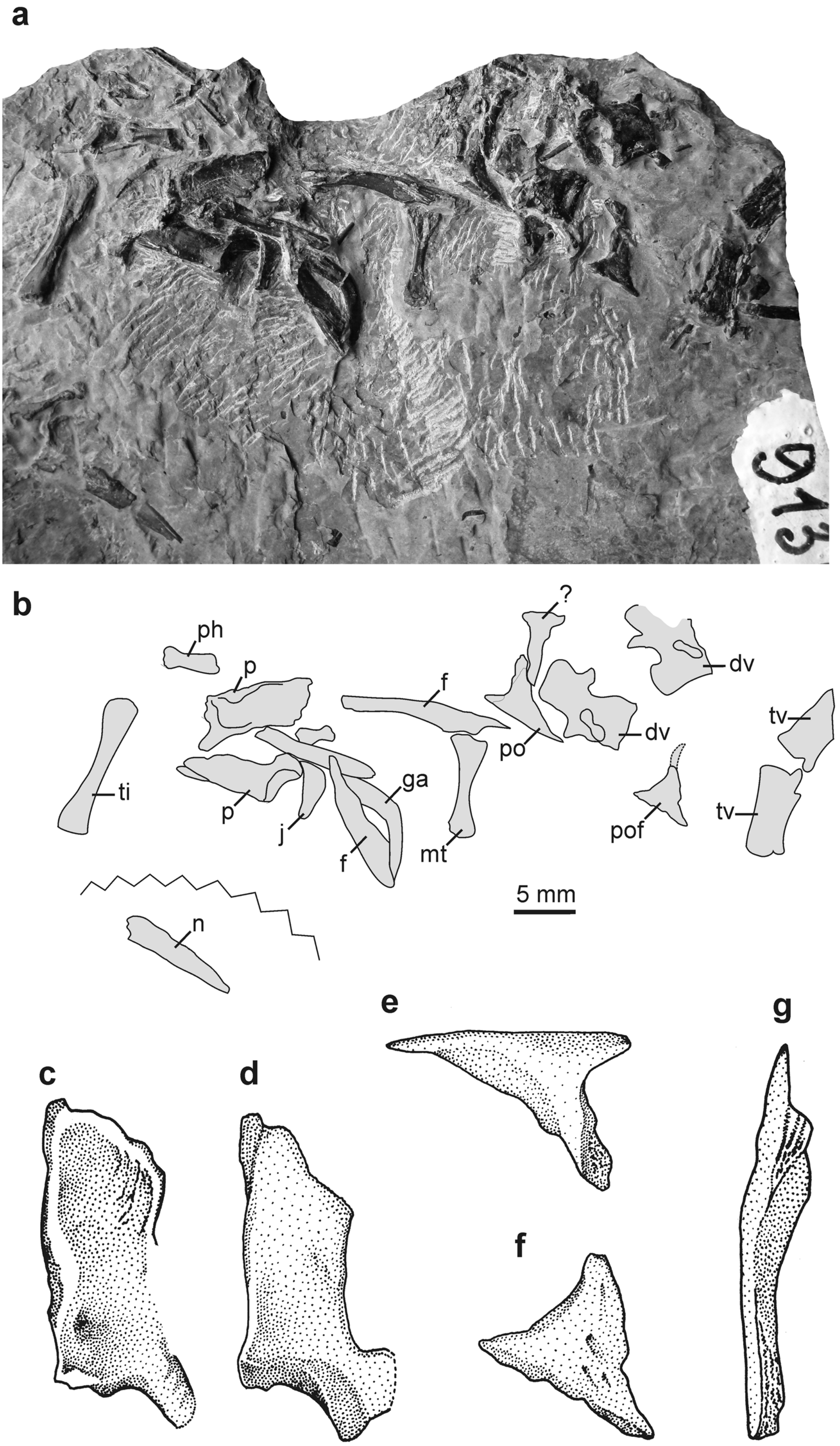
1. Gilbert, S. F., Loredo, G. A., Brukman, A. & Burke, A. C. Morphogenesis of the turtle shell: the development of a novel structure in tetrapod evolution. *Evol. Dev.* **3,** 47–58 (2001).
2. Rieppel, O. & Reisz, R. R. The origin and early evolution of turtles. *Annu. Rev. Ecol. Syst.* **30,** 1–22 (1999).
3. Gilbert, S. F., Bender, G., Betters, E., Yin, M. & Cebra-Thomas, J. A. The contribution of neural crest cells to the nuchal bone and plastron of the turtle shell. *Integr. Comp. Biol.* **47,** 401–408 (2007).
4. Lyson, T. R., Bever, G. S., Bhullar, B.-A. S., Joyce, W. G. & Gauthier, J. A. Transitional fossils and the origin of turtles. *Biol. Lett.* **6,** 830–833 (2010).
5. Lyson, T. R., Bever, G. S., Scheyer, T. M., Hsiang, A. Y. & Gauthier, J. A. Evolutionary origin of the turtle shell. *Curr. Biol.* **23,** 1113–1119 (2013).
6. Rieppel, O. in *Morphology and Evolution of Turtles* (eds Brinkman, D. B., Holroyd, P. A. & Gardner, J. D.) 51–61 (Springer, 2013).
7. Hedges, S. B. & Poling, L. L. A molecular phylogeny of reptiles. *Science* **283,** 998–1001 (1999).
8. Crawford, N. G. *et al.* More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biol. Lett.* **8,** 783–786 (2012).
9. Lee, M. S. Y. Turtle origins: insights from phylogenetic retrofitting and molecular scaffolds. *J. Evol. Biol.* **26,** 2729–2738 (2013).
10. Lu, B., Yang, W., Dai, Q. & Fu, J. Using genes as characters and a parsimony analysis to explore the phylogenetic position of turtles. *PLoS ONE* **8,** e79348 (2013).
11. Hirasawa, T., Pascual-Anaya, J., Kamezaki, N., Taniguchi, M., Mine, K. & Kuratani, S. The evolutionary origin of the turtle shell and its dependence on the axial arrest of the embryonic rib cage. *J. Exp. Zool. B* **324,** 194–207 (2015).
12. Li, C., Wu, X.-C., Rieppel, O., Wang, L.-T. & Zhao, L.-J. An ancestral turtle from the Late Triassic of southwestern China. *Nature* **456,** 497–501 (2008).
13. Zangerl, R. The homology of the shell elements in turtles. *J. Morphol.* **65,** 383–406 (1939).
14. deBraga, M. & Rieppel, O. Reptile phylogeny and the affinities of turtles. *Zool. J. Linn. Soc.* **120,** 281–354 (1997).
15. Müller, J. in *Recent Advances in the Origin and Early Radiation of Vertebrates* (eds Arratia, G., Wilson, M. V. H. & Cloutier, R.) 379–408 (Dr Friedrich Pfeil, 2004).
16. Joyce, W. G., Parham, J. F. & Gauthier, J. A. Developing a protocol for the conversion of rank-based taxon names to phylogenetically defined clade names, as exemplified by turtles. *J. Paleontol.* **78,** 989–1013 (2004).
17. Schoch, R. R. Stratigraphie und Taphonomie wirbeltierreicher Schichten im Unterkeuper (Mitteltrias) von Vellberg (SW-Deutschland). *Stuttgart. Beitr. Naturk. B* **318,** 1–30 (2002).
18. Gaffney, E. S. The comparative osteology of the Triassic turtle *Proganochelys. Bull. Am. Mus. Nat. Hist.* **194,** 1–263 (1990).
19. Watson, D. M. S. *Eunotosaurus africanus* Seeley, and the ancestry of the Chelonia. *Proc. Zool. Soc. Lond.* **1914,** 1011–1020 (1914).
20. Li, C., Yang, D.-Y., Cheng, L., Wu, X.-C. & Rieppel, O. A new species of *Largocephalosaurus* (Diapsida: Saurosphargidae), with implications for the morphological diversity and phylogeny of the group. *Geol. Mag.* **151,** 100–120 (2014).
21. Hirasawa, T., Nagashima, H. & Kuratani, S. The endoskeletal origin of the turtle carapace. *Nat. Commun.* **4,** 2107 (2013).
22. Gow, C. E. A reassessment of *Eunotosaurus africanus* Seeley (Amniota: Parareptilia). *Palaeont. Afr.* **34,** 33–42 (1997).
23. Bever, G. S., Lyson, T. & Bhullar, B.-A. Fossil evidence for a diapsid origin of the anapsid turtle skull. *Soc. Vert. Paleont. Abstr.* **2014,** 91 (2014).
24. Joyce, W. G. & Gauthier, J. A. Palaeoecology of Triassic stem turtles sheds new light on turtle origins. *Proc. R. Soc. Lond. B* **271,** 1–5 (2003).
25. Scheyer, T. M. & Sander, P. M. Shell bone histology indicates terrestrial palaeoecology of basal turtles. *Proc. R. Soc. B* **274,** 1885–1893 (2007).

Supplementary Information is available in the online version of the paper.

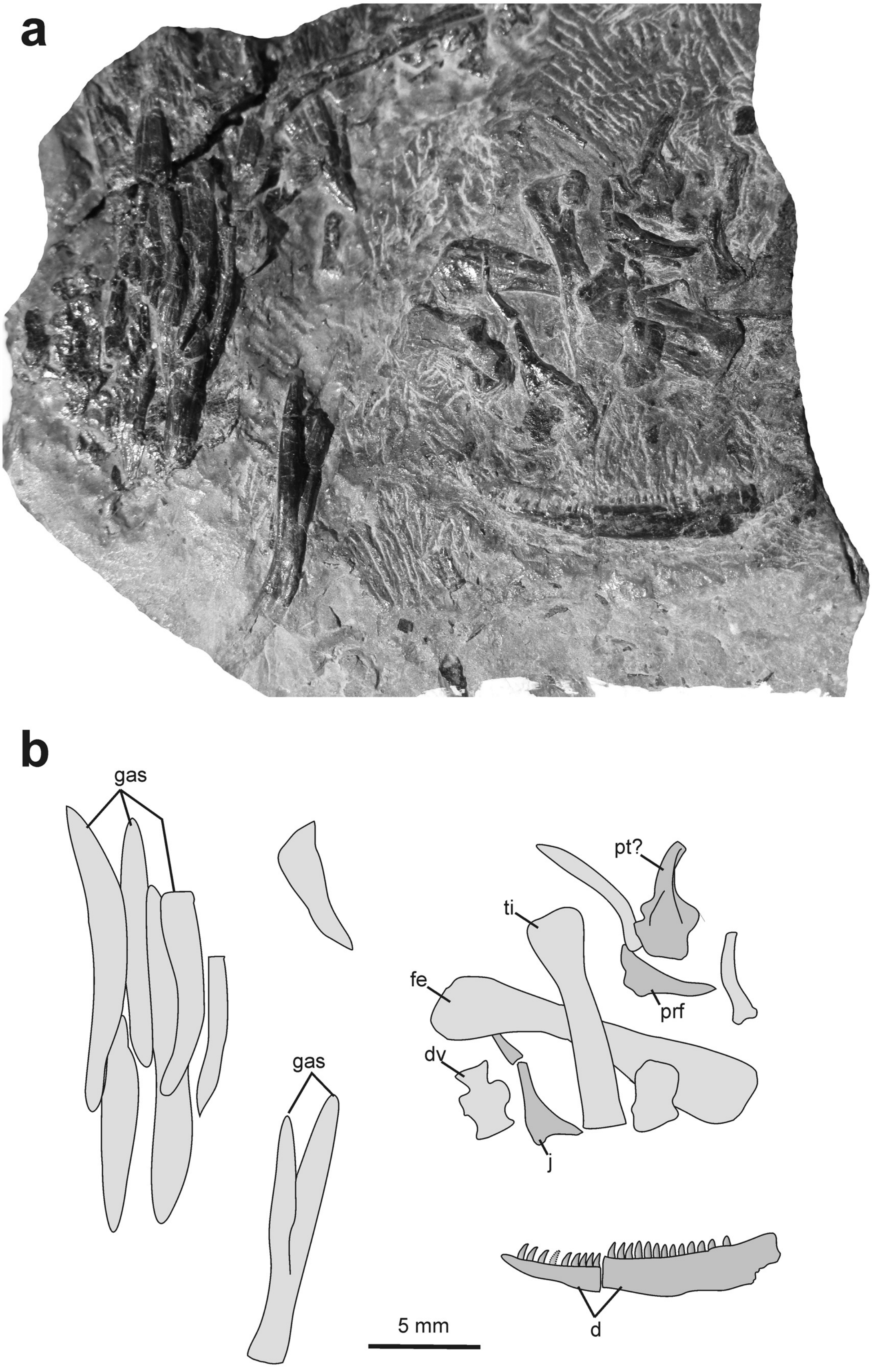
Acknowledgements We acknowledge the Schumann family for their continued support and granting access to the Vellberg quarry, and F. Ullmann, B. Rozynek, W. Kugler, T. Haubold, U. Günter, and M. Salomon for assistance in the field and for donating specimens to the Staatliches Museum für Naturkunde Stuttgart. N. Klein and T. Scheyer assisted with the histological interpretation of the thin-sections. T. Lyson provided the character-taxon matrix used in ref. 5. I. Rosin, N. Adorf, M. Kamenz, and K. Krämer prepared the material, and C. Wimmer-Pfeil prepared thin-sections. We thank D. Seegis, H. Hagdorn, W. Joyce, N. Klein, T. Lyson, J. Müller, O. Rieppel, and T. Scheyer for discussions.

Author Contributions R.R.S. and H.-D.S. contributed equally to the research and the development of the manuscript; therefore their names are listed in alphabetical order.

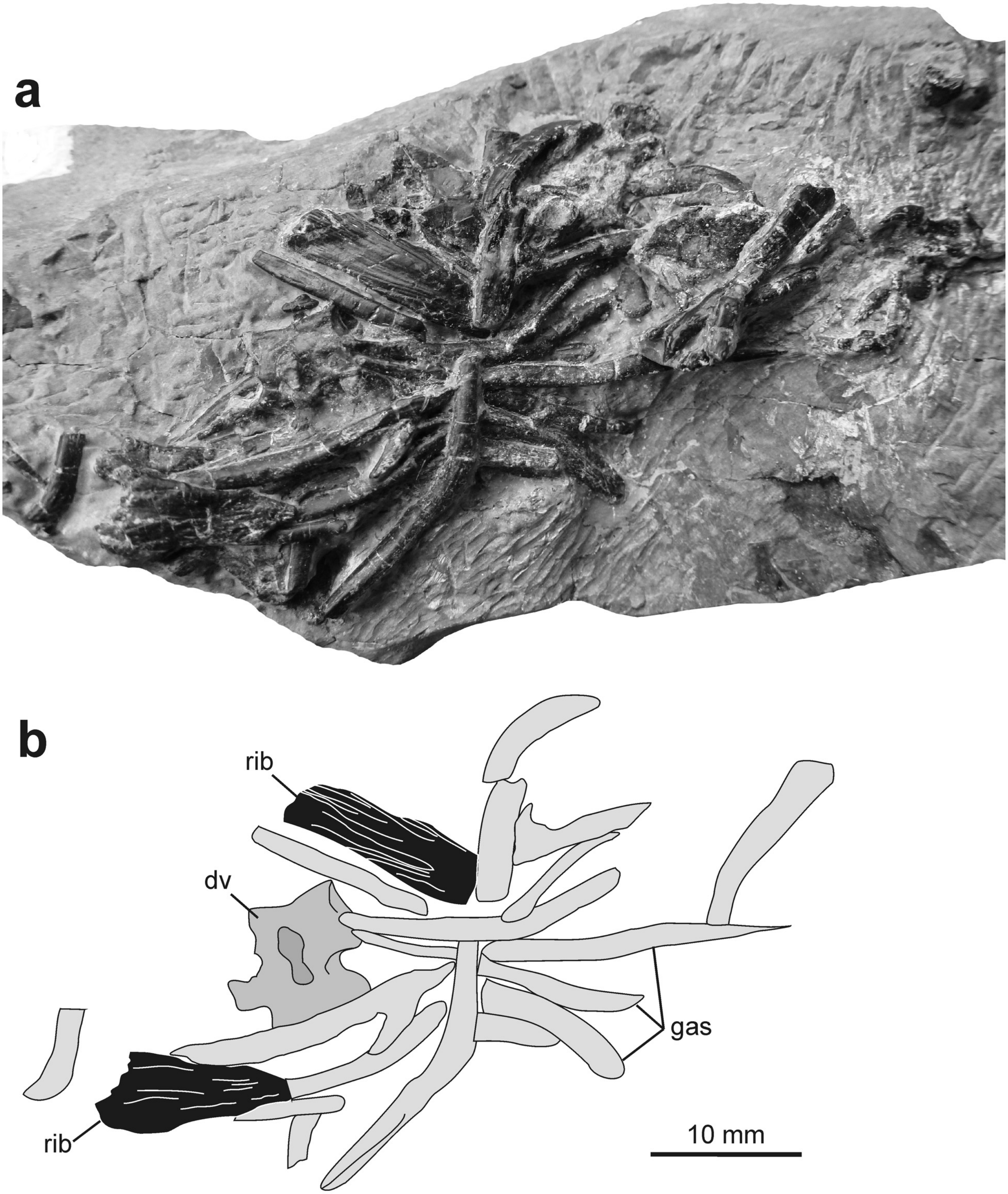
Author Information *P. rosinae* is in the ZooBank database (http://zoobank.org/) with Life Science Identifier urn:lsid:zoobank.org:act:CDD54976-047F-43AA-80F4-9680DF78CD7B. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.R.S. (rainer.schoch@smns-bw.de) or H.-D.S. (suesh@si.edu).


a
913
b
ph
p
f
?
dv
tv
ti
p
j
ga
f
mt
po
dv
pof
tv
n
5 mm
c
d
e
f
g

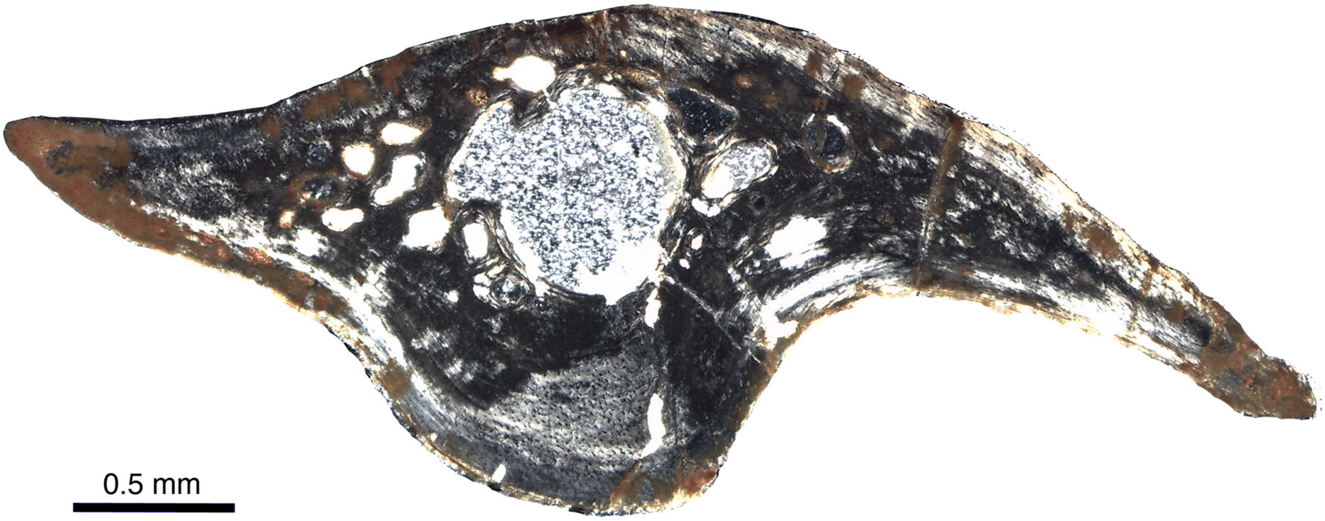
Extended Data Figure 1 | Cranial material of *P. rosinae*. **a**, **b**, Photograph and explanatory outline drawing of partial skull and postcranial skeleton of *P. rosinae* (SMNS 91356); **c**, left parietal in ventral view; **d**, right parietal in dorsal view; **e**, left postorbital; **f**, left postfrontal; **g**, left frontal. Abbreviations: dv, dorsal vertebra; f, frontal; ga, gastralium; j, jugal; mt, metatarsal; n, nasal; p, parietal; ph, phalanx; po, postorbital; pof, postfrontal; ti, tibia; tv, tail vertebra.



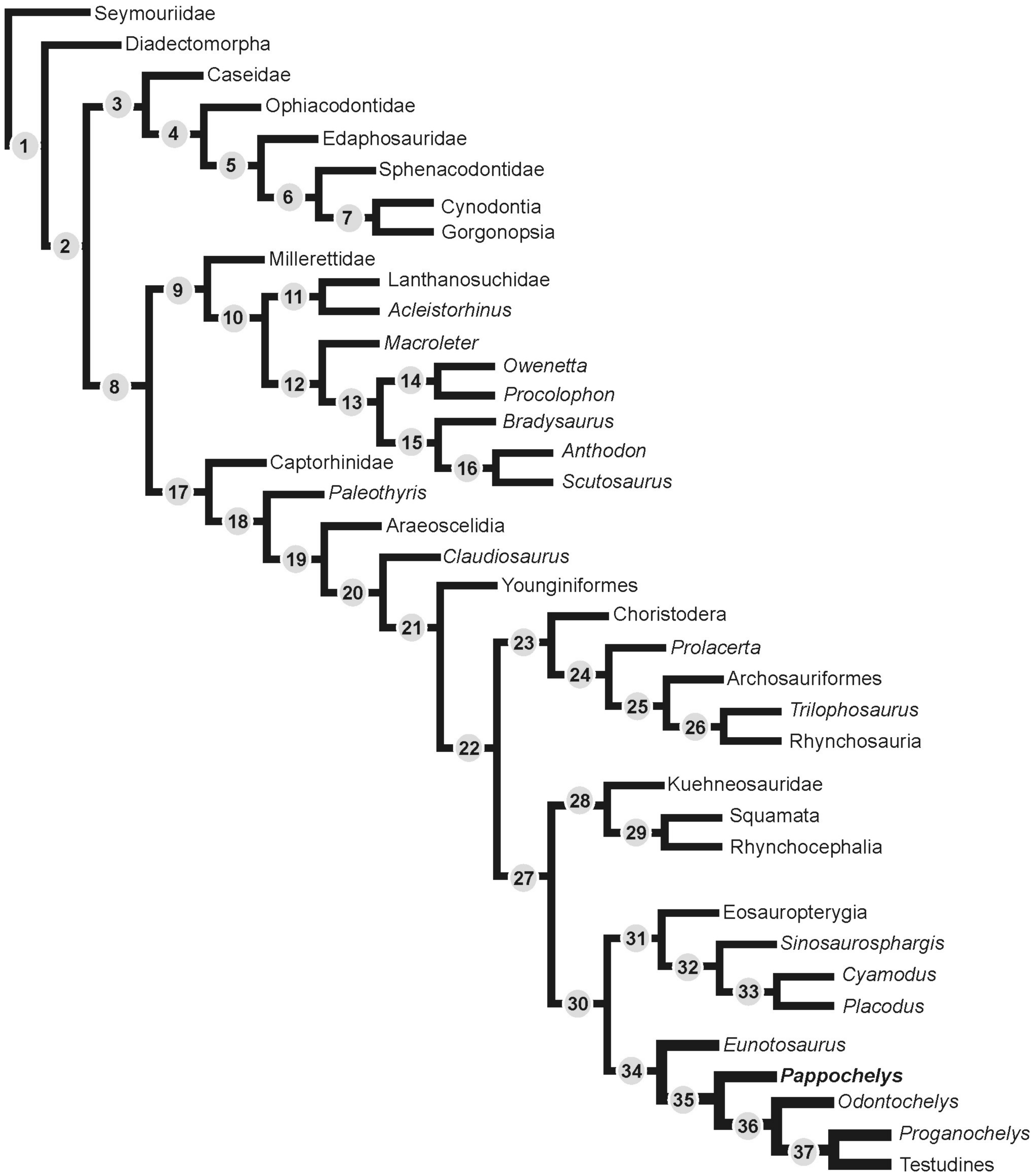
Extended Data Figure 2 | Skeletal remains of a very small individual of *P. rosinae.* **a**, **b**, Photograph (**a**) and explanatory outline drawing (**b**) of associated skeletal remains of a very small individual of *P. rosinae* (SMNS 92066). Bones of the skull are shown in a darker shade of grey. Abbreviations: d, dentary; dv, dorsal vertebra; fe, femur; gas, gastralia; j, jugal; prf, prefrontal; pt?, possible pterygoid; ti, tibia.



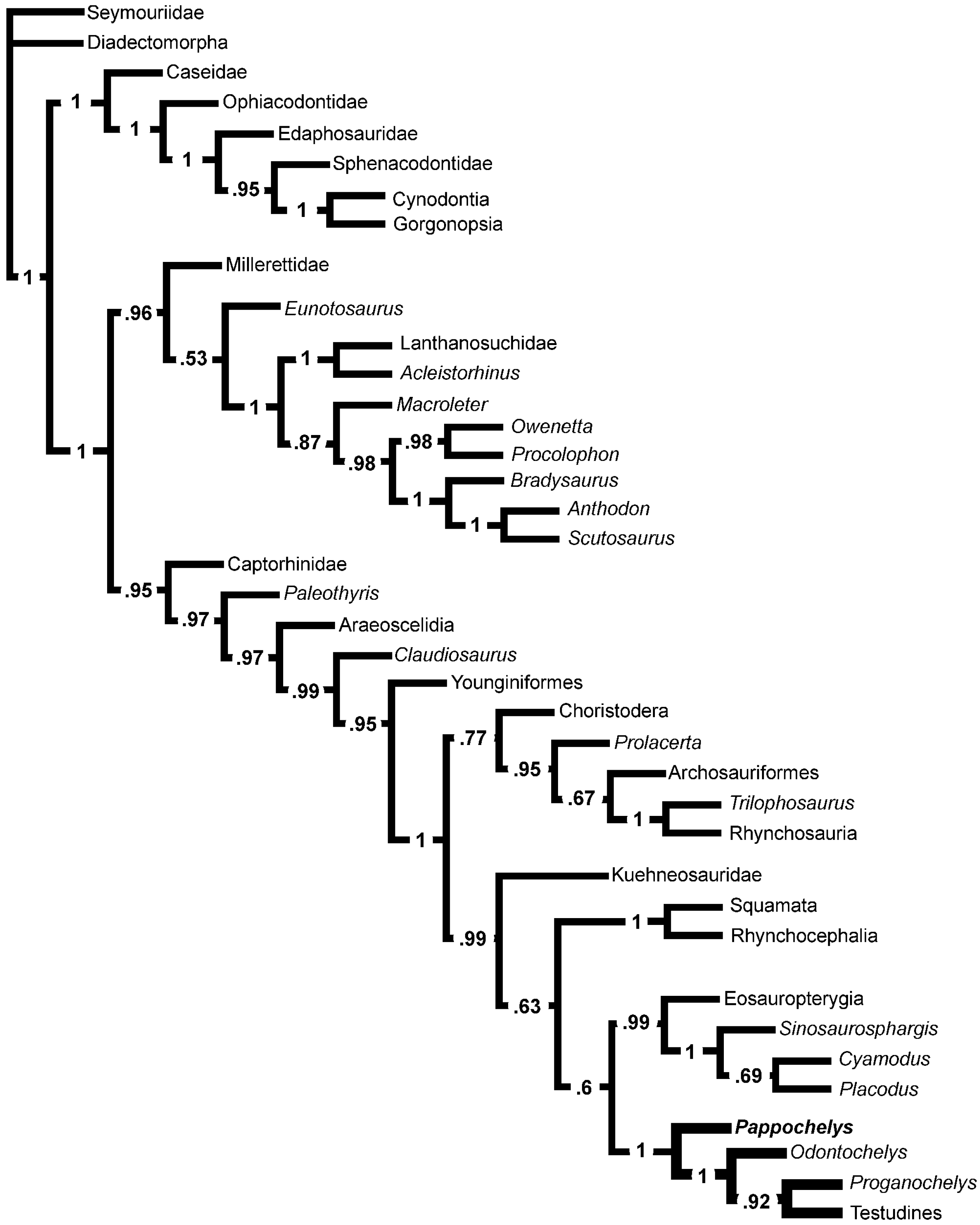
Extended Data Figure 3 | Gastralia of *P. rosinae*. **a**, **b**, Photograph (**a**) and explanatory outline (**b**) of a set of gastralia elements and fragments of two trunk ribs (black) that are part of the incomplete, partly articulated postcranial skeleton SMNS 91360.



Extended Data Figure 4 | Transverse section through the broadened shaft of a left trunk rib.



Extended Data Figure 5 | Tree illustrating hypothesis of turtle relationships based on the Tree Analysis using New Technology (TNT) program. Individual nodes are numbered. For additional information refer to 'Phylogenetic analysis' section in Supplementary Information.



Extended Data Figure 6 | Tree illustrating hypothesis of turtle relationships based on Bayesian analysis. Numbers at individual nodes represent posterior probabilities. For additional information refer to 'Phylogenetic analysis' section in Supplementary Information.

doi:10.1038/nature14659

Sparse whole-genome sequencing identifies two loci for major depressive disorder

CONVERGE consortium*

Major depressive disorder (MDD), one of the most frequently encountered forms of mental illness and a leading cause of disability worldwide[1], poses a major challenge to genetic analysis. To date, no robustly replicated genetic loci have been identified[2], despite analysis of more than 9,000 cases[3]. Here, using low-coverage whole-genome sequencing of 5,303 Chinese women with recurrent MDD selected to reduce phenotypic heterogeneity, and 5,337 controls screened to exclude MDD, we identified, and subsequently replicated in an independent sample, two loci contributing to risk of MDD on chromosome 10: one near the *SIRT1* gene ($P = 2.53 \times 10^{-10}$), the other in an intron of the *LHPP* gene ($P = 6.45 \times 10^{-12}$). Analysis of 4,509 cases with a severe subtype of MDD, melancholia, yielded an increased genetic signal at the *SIRT1* locus. We attribute our success to the recruitment of relatively homogeneous cases with severe illness.

The existence and number of subtypes of depression have been debated over the past 100 years. The current consensus is that depression may be a collection of partly distinct diseases, with overlapping causal pathways. This aetiologic heterogeneity might therefore substantially reduce the power of genetic association studies, and hence explain the failure to find genetic risk loci[3]. For example, there may be cases of MDD of largely environmental origin whose presence reduces the power to detect genetic effects. Also, genetic risk factors for mild depressive syndromes may not be entirely the same as those for more severe cases[4].

For these reasons, we investigated the genetic basis of MDD in subjects for whom known sources of phenotypic and genetic heterogeneity were minimized and known risk factors documented. The CONVERGE (China, Oxford and Virginia Commonwealth University Experimental Research on Genetic Epidemiology) consortium recruited 11,670 Han Chinese women through a collaboration involving 58 hospitals in China. We studied only women because about 45% of the genetic liability to MDD is not shared between sexes[5,6]. In an attempt to obtain severe cases of MDD, we recruited only recurrent cases (mean number of episodes was 5.6).

We used low-coverage sequencing to genotype our sample[7]. Whole-genome sequences were acquired to a mean depth of 1.7× (95% confidence intervals (CIs) 0.7–4.3) per individual, from which 32,781,340 SNP sites were identified. After applying stringent quality controls (Methods), we obtained 10,640 samples (5,303 cases of MDD, 5,337 controls) and 6,242,619 SNPs for inclusion in genome-wide association studies (GWAS). We compared genotypes from the low-coverage sequencing to genotypes called with 10× coverage sequence and to genotypes called from genotyping arrays and a mass spectrometer platform. The mean percentage concordance between genotypes from nine individuals with both low- and 10× coverage across all sites was 98.1% (Supplementary Table 1). We compared imputed genotypes to those acquired for 72 individuals using an array and to 21 SNPs genotyped on all individuals with the MassARRAY system mass spectrometer (Supplementary Notes). Overall concordance was 98.0% (Supplementary Tables 2 and 3).

Genetic association analysis was carried out with a linear mixed model with a genetic relatedness matrix (GRM) as a random effect and principal components from eigen-decomposition of the GRM as fixed effect covariates (Methods, Supplementary Notes)[8,9]. Fig. 1a and Extended Data Fig. 1 show the Manhattan and quantile–quantile plots, respectively, for this analysis. The genomic control inflation factor (λ, the ratio of the observed median χ^2 to that expected by chance) for association with MDD was 1.070 (for common SNPs, minor allele frequency (MAF) $>2\%$, $\lambda = 1.074$). The adjusted measure for sample size to that of 1,000 cases and 1,000 controls (λ_{1000}) was 1.013.

Two loci exceeded genome-wide significance in association with MDD: one 5′ to the sirtuin1 (*SIRT1*) gene on chromosome 10 (SNP = rs12415800, chromosome 10:69624180, MAF = 45.2%, $P = 1.92 \times 10^{-8}$, Fig. 1b), and the other in an intron of the phospholysine phosphohistidine inorganic pyrophosphate phosphatase (*LHPP*) gene (SNP = rs35936514, chromosome 10:126244970, MAF = 26.0%, $P = 1.27 \times 10^{-8}$, Fig. 1c). All SNPs with P values of association $<10^{-5}$ with MDD are listed in Supplementary Table 4.

We checked the accuracy of the imputed genotypes at 12 SNPs with $P < 1 \times 10^{-5}$, by re-genotyping the CONVERGE samples using a MassARRAY system mass spectrometer, thereby confirming their association with MDD. Extended Data Table 1 shows that the correlation between the two assays was high (mean $r^2 = 0.984$), and the odds ratios for the two genome-wide significant SNPs assessed by the two methods were almost identical, with highly overlapping confidence intervals (rs12415800 odds ratios: 1.167 versus 1.167; rs35936514 odds ratios: 0.845 versus 0.842).

We replicated the associations by genotyping the same 12 SNPs in a separate Han Chinese cohort of 3,231 cases with recurrent MDD, and 3,186 controls (both sexes). Two SNPs at the peaks of association for *SIRT1* and *LHPP* loci (rs12415800 and rs35936514, respectively) for MDD in the CONVERGE samples were significantly associated with MDD (Table 1). Analysis of the combined samples gave P values for association with MDD at these two SNPs of 2.53×10^{-10} and 6.45×10^{-12}, respectively. Extended Data Table 2 shows the genotype distribution and P values for tests of violation of the Hardy–Weinberg equilibrium in both the CONVERGE samples and the replication cohort at both SNPs.

Comparison with results from the Psychiatric Genomics Consortium (PGC) mega-analysis of European studies[3] failed to provide robust replication for our top SNPs (Extended Data Fig. 2 and Extended Data Table 3). However, the proportion of associations in the same direction in the two studies exceeded expectations due to chance ($P < 0.001$), and polygenic risk scores from the PGC mega-analysis applied to the CONVERGE samples were of significant ($P < 0.01$) but limited predictive value, accounting for 0.1% of MDD risk in the CONVERGE cohort (Extended Data Table 4). It is unclear to what extent differences in sample ascertainment, ethnicity, or other factors contribute to the failure to replicate genetic effects in the PGC sample. Notably, variants at our most strongly associated loci are much rarer in European populations, where rs12415800 (*SIRT1*)

*A full list of authors and affiliations appears at the end of the paper.

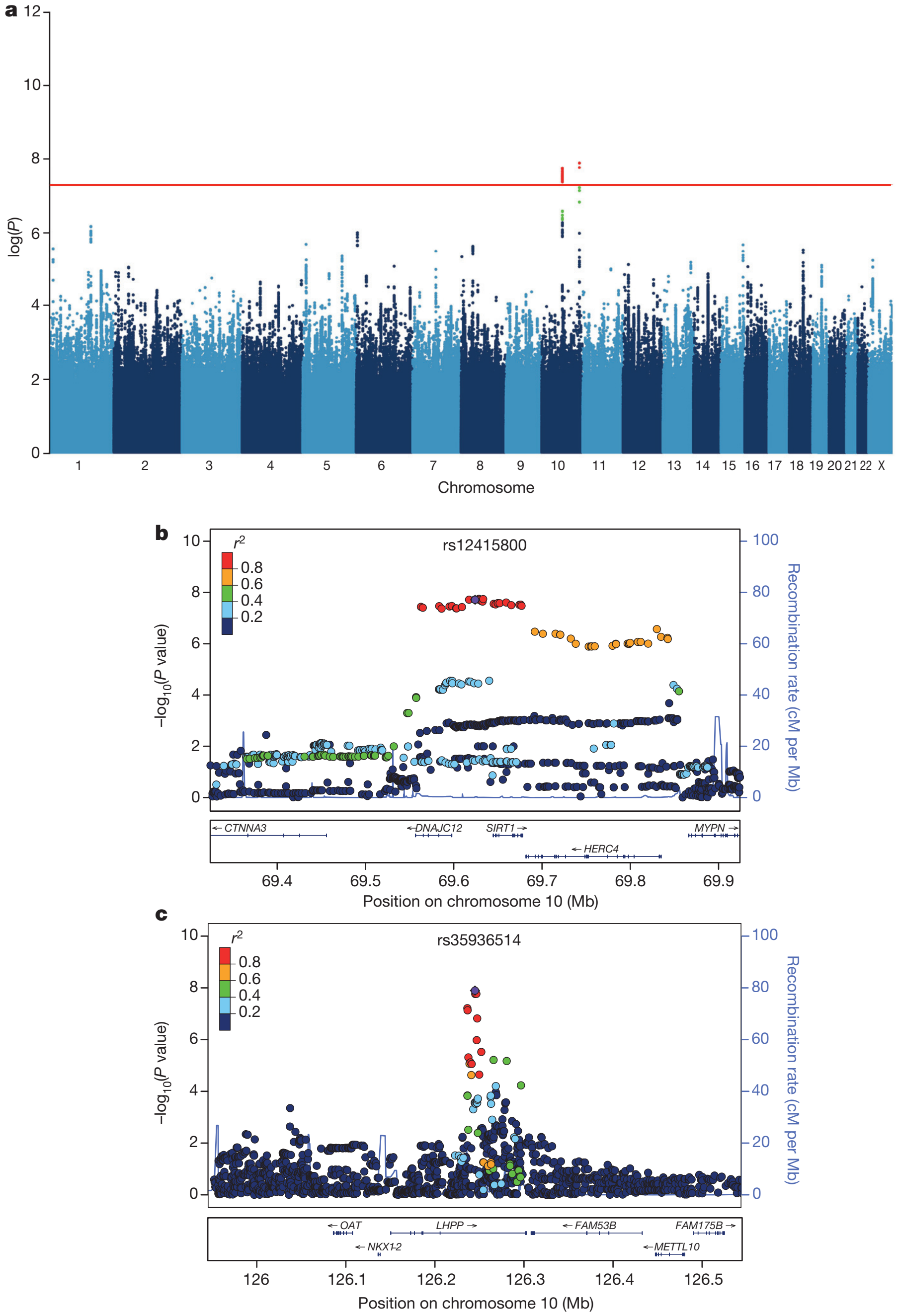


Figure 1 | Two loci associated with MDD in the CONVERGE sample. **a**, Manhattan plot of genome-wide association for MDD. **b**, Association at the *SIRT1* region on chromosome 10 at 69.6 megabases (Mb). **c**, Association at the *LHPP* gene on chromosome 10 at position 126.2 Mb. For **b** and **c** The $-\log_{10}(P \text{ value})$ of imputed SNPs associated with MDD is shown on the left *y* axis. The recombination rates expressed in centimorgans (cM) per Mb (NCBI Build GRCh37; light blue lines), are shown on the right *y* axis. Position in Mb is on the *x* axis. Linkage disequilibrium of each SNP with the top SNP, displayed as a large purple diamond, is indicated by its colour. The plots were drawn using LocusZoom[20].

and rs35936514 (*LHPP*) have frequencies of 3% and 8% respectively, compared to 45% and 26% in the CONVERGE cohort.

We considered whether successful mapping of MDD in the CONVERGE samples was attributable to the recruitment of a severe, more genetically determined form of the disease. We tested that hypothesis by looking within the CONVERGE cohort at a particularly severe, and more heritable form of MDD: melancholia[10]. Prior research has suggested that MDD patients with melancholia have more impairing, recurrent episodes and that risk for MDD is higher in the co-twins of probands with the melancholic subtype[11] than in those with non-melancholic MDD. This increase is greater in monozygotic than dizygotic twin pairs[11], as would be expected if the subtype were associated with greater genetic risk.

In the CONVERGE cohort, 85% of cases met the DSM-IV criteria for melancholia[12]. We searched for a genetic association in 9,846 samples (4,509 cases and 5,337 controls) and identified the same two loci that exceeded genome-wide significance on chromosome 10. The genomic control inflation factor λ for melancholia was 1.069, and λ_{1000} was 1.014. Even though the sample for melancholia was smaller than for MDD, at the *SIRT1* locus the significance of association was two orders of magnitude greater than for MDD (top SNP = rs80309727, chromosome 10:69617347, MAF = 45.2%, $P = 2.95 \times 10^{-10}$). Extended Data Fig. 3 shows the Manhattan plot, quantile–quantile plot and detailed views of the *SIRT1* locus associated with melancholia. All SNPs with *P* values of association $<10^{-5}$ with melancholia are listed in Supplementary Table 5. To determine whether the increased association might have arisen by chance, we generated an empirical distribution of odds ratios by randomly selecting 4,509 cases from the total set and re-analysing the association with each of the genome-wide significant variants. We found that the observed value lay on the 98.8th percentile at the *SIRT1* locus, but at the 61.6th percentile at the *LHPP* locus (Extended Data Fig. 4).

Our results indicate that, as others have suggested[13], obtaining low-sequence coverage of a large number of individuals can be an effective

Table 1 | Genetic association between MDD and 12 variants in the CONVERGE cohort and a replication sample

SNP					CONVERGE ($n = 10,640$)					Replication ($n = 6,417$)			Joint ($n = 17,057$)		
Chr.	Pos.	RSID	Ref.	Alt.	Freq.	Info.	OR	s.e.	*P*	OR	s.e.	*P*	OR	s.e.	*P*
1	11493832	rs2922240	T	C	0.385	1.018	1.141	0.028	2.80×10^{-6}	0.949	0.037	1.54×10^{-1}	1.070	0.022	2.46×10^{-3}
1	175151950	rs3766688	T	C	0.394	1.003	0.875	0.028	1.83×10^{-6}	0.991	0.037	8.15×10^{-1}	0.918	0.022	1.34×10^{-4}
1	228052027	rs57047840	A	G	0.284	0.970	1.138	0.031	4.64×10^{-5}	1.001	0.041	9.90×10^{-1}	1.088	0.025	5.57×10^{-4}
5	9161674	rs55713588	A	G	0.096	0.893	1.278	0.050	6.04×10^{-7}	1.054	0.062	3.93×10^{-1}	1.042	0.035	2.08×10^{-1}
6	4386107	rs55800092	C	T	0.151	1.001	0.824	0.039	1.35×10^{-6}	0.962	0.052	4.49×10^{-1}	0.876	0.031	1.82×10^{-5}
10	**69624180**	**rs12415800**	**G**	**A**	**0.452**	**0.992**	**1.164**	**0.028**	$\mathbf{1.92 \times 10^{-8}}$	**1.130**	**0.036**	$\mathbf{7.71 \times 10^{-4}}$	**1.150**	**0.022**	$\mathbf{2.37 \times 10^{-10}}$
10	**126244970**	**rs35936514**	**C**	**T**	**0.260**	**0.993**	**0.839**	**0.032**	$\mathbf{1.27 \times 10^{-8}}$	**0.838**	**0.041**	$\mathbf{1.68 \times 10^{-5}}$	**0.842**	**0.025**	$\mathbf{6.43 \times 10^{-12}}$
13	107659212	rs61967003	C	T	0.017	0.999	1.645	0.109	6.70×10^{-6}	0.788	0.150	1.11×10^{-1}	1.277	0.087	4.81×10^{-3}
14	66833851	rs17827252	C	G	0.463	1.011	0.887	0.028	1.44×10^{-5}	0.962	0.041	3.41×10^{-1}	0.907	0.023	2.20×10^{-5}
19	34493757	rs11880240	C	G	0.068	1.019	1.291	0.055	8.02×10^{-6}	1.048	0.072	5.12×10^{-1}	1.184	0.043	9.15×10^{-5}
X	24656658	rs1921918	A	G	0.721	0.995	0.883	0.031	3.22×10^{-5}	0.994	0.047	9.01×10^{-1}	0.917	0.026	1.09×10^{-3}
X	25011374	rs11573525	C	T	0.260	0.971	1.160	0.032	5.86×10^{-6}	1.011	0.047	8.19×10^{-1}	1.100	0.027	2.18×10^{-4}

The table reports results for 12 SNPs in the CONVERGE and replication samples. The first five columns give the chromosome (Chr.), genomic position (Pos.), SNP identifier (RSID), reference allele (Ref.) on Human Genome Reference GRCh37.p5 and alternative allele (Alt.) called in CONVERGE. The next five columns show the alternative allele frequency (Freq.) and results of association testing with MDD using imputed allele dosages in 10,640 CONVERGE samples (5,303 cases, 5,337 controls); information scores (Info.), odds ratio (OR) of association with MDD with respect to the alternative allele and standard error (s.e.) in the odds ratio were obtained from a logistic regression model; *P* values of association (*P*) were obtained from a linear-mixed model with a GRM containing all samples. The next three columns present the results of association with MDD in the replication cohort of 6,417 samples (3,231 cases, 3,186 controls) from a logistic regression model. The final three columns present the results of association with MDD in a joint analysis with both CONVERGE and replication cohorts from a logistic regression model. Bold type indicates the genome-wide significant markers.

way to screen the genome for association signals. We were able to genotype more variants than on genotyping arrays and our set is larger than publicly available sources for imputation[14]. Our imputation pipeline employed standard tools, and it is likely that imputation accuracy could be improved with further algorithmic research.

MDD is most probably highly polygenic[3], and many additional loci remain to be discovered. We attribute the discovery and replication of two SNPs associated with MDD in the CONVERGE cohort to the recruitment of cases who were probably more homogeneous and more severely impaired than those collected in previous studies from Western cultures. In East Asia, reluctance to report MDD[15] probably explains why hospital-ascertained cases are more severe, and why prevalence estimates for MDD are lower in China (3.6%[16]) than in the US (16.2%)[17]. Consistent with this interpretation, 85% of the cases of MDD in the CONVERGE cohort have melancholia, a severe subtype of MDD; mapping melancholia led to a significant increase in the genetic signal at one locus. Finally, we note that one of the replicated risk loci is located close to a gene involved in mitochondrial biogenesis (*SIRT1*)[18], which, together with our finding that MDD is associated with increased amounts of mitochondrial DNA[19], suggests an unexpected origin for at least some of the phenotypic manifestations of MDD.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 December 2014; accepted 11 June 2015.

Published online 15 July 2015.

1. Kessler, R. C. & Bromet, E. J. The epidemiology of depression across cultures. *Annu. Rev. Public Health* **34,** 119–138 (2013).
2. Flint, J. & Kendler, K. S. The genetics of major depression. *Neuron* **81,** 484–503 (2014).
3. Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium *et al.* A mega-analysis of genome-wide association studies for major depressive disorder. *Mol. Psychiatry* **18,** 497–511 (2013).
4. Foley, D. L. *et al.* Genetic and environmental risk factors for depression assessed by subject-rated symptom check list versus structured clinical interview. *Psychol. Med.* **31,** 1413–1423 (2001).
5. Kendler, K. S. *et al.* Clinical indices of familial depression in the Swedish Twin Registry. *Acta Psychiatr. Scand.* **115,** 214–220 (2007).
6. Sullivan, P. F. *et al.* Genetic epidemiology of major depression: review and meta-analysis. *Am. J. Psychiatry* **157,** 1552–1562 (2000).
7. Li, Y. *et al.* Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* **21,** 940–951 (2011).
8. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* **8,** 833–835 (2011).
9. Widmer, C. *et al.* Further improvements to linear mixed models for genome-wide association studies. *Sci. Rep.* **4,** 6874 (2014).
10. Angst, J. *et al.* Melancholia and atypical depression in the Zurich study: epidemiology, clinical characteristics, course, comorbidity and personality. *Acta Psychiatr. Scand.* Suppl. **433,** 72–84 (2007).
11. Kendler, K. S. The diagnostic validity of melancholic major depression in a population-based sample of female twins. *Arch. Gen. Psychiatry* **54,** 299–304 (1997).
12. Sun, N. *et al.* A comparison of melancholic and nonmelancholic recurrent major depression in Han Chinese women. *Depress. Anxiety* **29,** 4–9 (2012).
13. Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genet.* **44,** 631–635 (2012).
14. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
15. Liao, S. C. *et al.* Low prevalence of major depressive disorder in Taiwanese adults: possible explanations and implications. *Psychol. Med.* **42,** 1227–1237 (2012).
16. Lee, S. *et al.* The epidemiology of depression in metropolitan China. *Psychol. Med.* **39,** 735–747 (2009).
17. Kessler, R. C. *et al.* The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *J. Am. Med. Assoc.* **289,** 3095–3105 (2003).
18. Gerhart-Hines, Z. *et al.* Metabolic control of muscle mitochondrial function and fatty acid oxidation through SIRT1/PGC-1α. *EMBO J.* **26,** 1913–1923 (2007).
19. Cai, N. *et al.* Molecular signatures of major depression. *Curr. Biol.* **25,** 1146–1156 (2015).
20. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26,** 2336–2337 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was funded by the Wellcome Trust (WT090532/Z/09/Z, WT083573/Z/07/Z, WT089269/Z/09/Z), NIH grant MH-100549 and the Brain and Behavior Research Foundation. All authors are part of the CONVERGE consortium (China, Oxford and VCU Experimental Research on Genetic Epidemiology) and gratefully acknowledge the support of all partners in hospitals across China. W. Kretzschmar is funded by the Wellcome Trust (WT097307). N. Cai is supported by the Agency of Science, Technology and Research (A*STAR) Graduate Academy. J. Marchini is funded by an ERC Consolidator Grant (617306). Q. Xu is funded by the 973 Program (2013CB531301) and NSFC (31430048, 31222031).

Author Contributions Manuscript preparation: N. Cai, T. B. Bigdeli, W. Kretzschmar, M. Reimers, T. Webb, B. Riley, S. Bacanu, R. E. Peterson, K. S. Kendler and J. Flint. Replication sample: Q. Xu CONVERGE sample collection: Yih. Li, Y. Chen, H. Deng, W. Sang, Ke. Li, J. Gao, B. Ha, S. Gao, J. Hu, C. Hu, G. Huang, G. Jiang, X. Zhou, You. Li, Kan Li, Q. Niu, Yi Li, G. Li, L. Liu, Z. Liu, Yi Li, X. Fang, R. Pan, G. Miao, Q. Zhang, F. Yu, G. Chen, M. Cai, D. Yang, X. Hong, Y. Song, C. Gao, J. Pan, Y. Zhang, T. Liu, J. Dong, X. Wang, L. Wang, Q. Mei, Z. Shen, X. Liu, W. Wu, D. Gu, Y. Chen, T. Liu, H. Rong, Yi. Liu, L. Lv, H. Meng, H. Sang, J. Shen, T. Tian, J. Shi, J. Sun, M. Tao, X. Wang, J. Xia, Q. He, G. Wang, X. Wang, Lina Yang, K. Zhang, N. Sun, J. Zhang, Z. Gan, Z. Zhang, W. Zhang, H. Zhong, F. Yang, E. Cong, S. Shi, G. Fu, J. Flint and K. S. Kendler. Genome sequencing and analysis: J. Liang, J. Hu, Q. Li, W. Jin, Z. Hu, G. Wang, Linm. Wang, P. Qian, Yu. Liu, T. Jiang, Y. Lu, X. Zhang, Y. Yin, Yin. Li, H. Yang, Jia. Wang, X. Gan, Yih. Li, N. Cai, R. Mott, J. Flint, Jun Wang and X. Xu. Genotype imputation: W. Kretzschmar, J. Hu, L. Song, Q. Li, N. Cai and J. Marchini. Genetic analysis: N. Cai, T. Bigdeli, Yih. Li, R. E. Peterson, S. Bacanu, T. Webb, B. Riley, K. S. Kendler, R. Mott and J. Flint.

Author Information All sequence data and MDD results are freely available at http://dx.doi.org/10.5524/100155. GWAS results are also available at http://www.med.unc.edu/pgc/downloads. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Q. Xu (xuqi@pumc.edu.cn), Jun Wang (wangj@genomics.org.cn), K. S. Kendler (kendler@vcu.edu) or J. Flint (jf@well.ox.ac.uk).

CONVERGE consortium

Na Cai[1]*, Tim B. Bigdeli[2]*, Warren Kretzschmar[1]*, Yihan Li[1]*, Jieqin Liang[3], Li Song[3], Jingchu Hu[3], Qibin Li[3], Wei Jin[3], Zhenfei Hu[3], Guangbiao Wang[3], Linmao Wang[3], Puyi Qian[3], Yuan Liu[3], Tao Jiang[3], Yao Lu[3], Xiuqing Zhang[3], Ye Yin[3], Yingrui Li[3], Xun Xu[3], Jingfang Gao[4], Mark Reimers[2], Todd Webb[2], Brien Riley[2], Silviu Bacanu[2], Roseann E. Peterson[2], Yiping Chen[5], Hui Zhong[6], Zhengrong Liu[7], Gang Wang[8], Jing Sun[9], Hong Sang[10], Guoqing Jiang[11], Xiaoyan Zhou[11], Yi Li[12], Yi Li[13], Wei Zhang[14], Xueyi Wang[15], Xiang Fang[16], Runde Pan[17], Guodong Miao[18], Qiwen Zhang[19], Jian Hu[20], Fengyu Yu[21], Bo Du[22], Wenhua Sang[22], Keqing Li[22], Guibing Chen[23], Min Cai[24], Lijun Yang[25], Donglin Yang[26], Baowei Ha[27], Xiaohong Hong[28], Hong Deng[29], Gongying Li[30], Kan Li[31], Yan Song[32], Shugui Gao[33], Jinbei Zhang[34], Zhaoyu Gan[34], Huaqing Meng[35], Jiyang Pan[36], Chengge Gao[37], Kerang Zhang[38], Ning Sun[38], Youhui Li[39], Qihui Niu[39], Yutang Zhang[40], Tieqiao Liu[41], Chunmei Hu[42], Zhen Zhang[43], Luxian Lv[44], Jicheng Dong[45], Xiaoping Wang[46], Ming Tao[47], Xumei Wang[48], Jing Xia[48], Han Rong[49], Qiang He[50], Tiebang Liu[51], Guoping Huang[52], Qiyi Mei[53], Zhenming Shen[54], Ying Liu[55], Jianhua Shen[56], Tian Tian[56], Xiaojuan Liu[57], Wenyuan Wu[58], Danhua Gu[59], Guangyi Fu[1], Jianguo Shi[60], Yunchun Chen[61], Xiangchao Gan[62], Lanfen Liu[63], Lina Wang[63], Fuzhong Yang[64], Enzhao Cong[64], Jonathan Marchini[1,65], Huanming Yang[3], Jian Wang[3], Shenxun Shi[64,66], Richard Mott[1], Qi Xu[67]§, Jun Wang[3,68,69,70]§, Kenneth S. Kendler[2]§ & Jonathan Flint[1,71]§

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. [2]Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, Virginia 23298, USA. [3]BGI-Shenzhen, Floor 9 Complex Building, Beishan Industrial Zone, YantianDistrict, Shenzhen, Guangdong 518083, China. [4]Zhejiang Traditional Chinese Medical Hospital, No.54 Youdian Road, Hangzhou, Zhejiang 310000, China. [5]CTSU, Richard Doll Building, University of Oxford, Old Road Campus, Oxford OX3 7LF, UK. [6]Anhui Mental Health Center, No.316 Huangshan Road, Hefei, Anhui 230000, China. [7]Anshan Psychiatric Rehabilitation Hospital, No.127 Shuangshan Road, Lishan District, Anshan, Liaoning 114000, China. [8]Beijing Anding Hospital of Capital University of Medical Sciences, No.5 Ankang Hutong, Deshengmen wai, Xicheng District, Beijing, Beijing 100000, China. [9]Brain Hospital of Nanjing Medical University, No.264 Guangzhou Road, Nanjing, Jiangsu 210000, China. [10]Changchun Mental Hospital, No.4596 Beihuan Road, Changchun, Jilin 130000, China. [11]Chongqing Mental Health Center, No.102 Jinzishan, Jiangbei District, Chongqing 404100, China. [12]Dalian No.7 Hospital, No.179 Lingshui Road, Ganjingzi District, Dalian, Liaoning 116000, China. [13]Wuhan Mental Health Center, No.70, Youyi Road, Wuhan, Hubei 430000, China. [14]Daqing No.3 Hospital of Heilongjiang Province, No.54 Xitai Road, Ranghulu district, Daqing, Heilongjiang 163000, China. [15]First Hospital of Hebei Medical University, No.89 Donggang Road, Shijiazhuang, Hebei 50000, China. [16]Fuzhou Psychiatric Hospital, No.451 South Erhuan Road, Cangshan District, Fuzhou, Fujian 350000, China. [17]Guangxi Longquanshan Hospital, No.1 Jila Road, Yufeng District, Liuzhou, Guangxi Zhuangzu 545000, China. [18]Guangzhou Brain Hospital, Guangzhou Psychiatric Hospital, No.36 Mingxin Road, Fangcun Avenue, Liwan District, Guangzhou, Guangdong 510000, China. [19]Hainan Anning Hospital, No.10 East Nanhai Avenue, Haikou, Hainan 570100, China. [20]Harbin Medical University, No.23 Youzheng street, Nangang District, Haerbin, Heilongjiang 150000, China. [21]Harbin No.1 Special Hospital, No.217 Hongwei Road, Haerbin, Heilongjiang 150000, China. [22]Hebei Mental Health Center, No.572 Dongfeng Road, Baoding, Hebei 71000, China. [23]Huaian No.3 Hospital, No.272 West Huaihai Road, Huaian, Jiangsu 223001, China. [24]Huzhou No.3 Hospital, No.255 Gongyuan Road, Huzhou, Zhejiang 313000, China. [25]Jilin Brain Hospital, No.98 West Zhongyang Road, Siping, Jilin 136000, China. [26]Jining Psychiatric Hospital, North Dai Zhuang, Rencheng District, Jining, Shandong 272000, China. [27]Liaocheng No. 4 Hospital, No.47 North Huayuan Road, Liaocheng, Shandong 252000, China. [28]Mental Health Center of Shantou University, No.243 Daxue Road, Shantou, Guangdong 515000, China. [29]Mental Health Center of West China Hospital of Sichuan University, No.28 South Dianxin Street, Wuhou District, Chengdu, Sichuan 610000, China. [30]Mental Health Institute of Jining Medical College, Dai Zhuang, Bei Jiao, Jining, Shandong 272000, China. [31]Mental Hospital of Jiangxi Province, No.43 Shangfang Road, Nanchang, Jiangxi 330000, China. [32]Mudanjiang Psychiatric Hospital of Heilongjiang Province, Xinglong, Mudanjiang, Heilongjiang 157000, China. [33]Ningbo Kang Ning Hospital, No.1 Zhuangyu Road, Zhenhai District, Ningbo, Zhejiang 315000, China. [34]No. 3 Hospital of Sun Yat-sen University, No.600 Tianhe Road, Tianhe District, Guangzhou, Guangdong 510630, China. [35]No.1 Hospital of Chongqing Medical University, No.1 Youyi Road,Yuanjiagang,Yuzhong District, Chongqing, Chongqing 400016, China. [36]No.1 Hospital of Jinan University, No.613 West Huangpu Avenue, Guangzhou, Guangdong 510000, China. [37]No.1 Hospital of Medical College of Xian Jiaotong University, No. 277 West Yan Ta Road, Xian, Shaan Xi 710061, China. [38]No.1 Hospital of Shanxi Medical University, No.85 South Jiefang Road, Taiyuan, Shanxi 30000, China. [39]No.1 Hospital of Zhengzhou University, No.1 East Jianshe Road, Zhengzhou, Henan 450000, China. [40]No.2 Hospital of Lanzhou University, No.82, Cuiyingmen, Lanzhou, Gansu 730000, China. [41]No.2 Xiangya Hospital of Zhongnan University, No.139 Middle Renmin Road, Furong District, Changsha, Hunan 410000, China. [42]No.3 Hospital of Heilongjiang Province, No.135 Jiaotong Road, Beian, Heilongjiang 164000, China. [43]No.4 Hospital of Jiangsu University, No.246 Nanmen Street, Zhenjiang, Jiangsu 212000, China. [44]Psychiatric Hospital of Henan Province, No.388 Middle Jianshe Road, Xinxiang, Henan 453000, China. [45]Qingdao Mental Health Center, No.299 Nanjing Road, Shibei District, Qingdao, Shandong 266000, China. [46]Renmin Hospital of Wuhan University, No.238 Jiefang Road, Wuchang District, Wuhan, Hubei 430000, China. [47]Second Affiliated Hospital of Zhejiang Chinese Medical University, No.318 Chaowang Road, Hangzhou, Zhejiang 310000, China. [48]ShengJing Hospital of China Medical University, No.36 Sanhao Street, Heping District, Shenyang, Liaoning 110001, China. [49]Shenzhen Key Lab for Psychological Healthcare, Kangning Hospital, No.1080, Cuizhu Street, Luohu District, Shenzhen, Guangdong 518000, China. [50]Department of General Internal Medicine, Kanazawa Medical University, Kahoku, Ishikawa 920-0293, Japan. [51]Shenzhen Key Lab for Psychological Healthcare;Shenzhen Kangning Hospital, No.1080, Cuizhu Street, Luohu District, Shenzhen, Guangdong 518000, China. [52]Sichuan Mental Health Center, No.190, East Jiannan Road, Mianyang, Sichuan 621000, China. [53]Suzhou Guangji Hospital, No.286, Guangji Road, Suzhou, Jiangsu 215000, China. [54]Tangshan No.5 Hospital, No.57 West Nanxin Road, Lunan District, Tangshan, Hebei 63000, China. [55]The First Hospital of China Medical University, No.155 North Nanjing Street, Heping District, Shenyang, Liaoning 110001, China. [56]Tianjin Anding Hospital, No.13 Liulin Road, Hexi District, Tianjin 300000, China. [57]Tianjin First Center Hospital, No.55 Xuetang Street, Xinkai Road, Hedong District, Tianjin 300000, China. [58]Tongji University Hospital, No.389 Xinchun Road, Shanghai 200000, China. [59]Weihai Mental Health Center, Qilu Avenue, ETDZ, Weihai, Shandong 264200, China. [60]Xian Mental Health Center, No.15 Yanyin Road, New Qujiang District, Xian, Shaanxi 710000, China. [61]Xijing Hospital of No.4 Military Medical University, No.17 West Changle Road, Xian, Shaanxi 710000, China. [62]Department of Comparative Developmental Genetics, Max Planck Institute for Plant Breeding Research, Carl-von-Linne-Weg 10, Cologne 50829, Germany. [63]Shandong Mental Health Center, No.49 East Wenhua Road, Jinan, Shandong 250000, China. [64]Shanghai Jiao Tong University School of Medicine, Shanghai Mental Health Centre, No. 600 Wan Ping Nan Road, Shanghai 200030, China. [65]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK. [66]Fudan University affiliated Huashan Hospital, No. 12 Wulumuqi Zhong Road, Shanghai 200040, China. [67]National Laboratory of Medical Molecular Biology, Institute of Basic Medical Sciences & Neuroscience Center, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 10005, China. [68]Department of Biology, University of Copenhagen, Ole Maal Oes Vej 5, Copenhagen 2200, Denmark. [69]Macau University of Science and Technology, Avenida Wai long, Taipa, Macau 999078, China, Taipa, Macau 999078, China. [70]Princess Al Jawhara Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah 21589, Saudi Arabia. [71]East China Normal University, 3663 North Zhongshan Road, Shanghai 200062, China.

*These authors contributed equally to this work.
§These authors jointly supervised this work.

METHODS

No statistical methods were used to predetermine sample size.

Sample collection. CONVERGE collected cases of recurrent major depression from 58 provincial mental health centres and psychiatric departments of general medical hospitals in 45 cities and 23 provinces of China. Controls were recruited from patients undergoing minor surgical procedures at general hospitals (37%) or from local community centres (63%). A sample size of 6,000 cases and 6,000 controls was chosen on the basis of evidence available when the study was designed (in 2007) of the likely existence of genetic loci with odds ratio of 1.2 and above. All subjects were Han Chinese women with four Han Chinese grandparents. Cases were excluded if they had a pre-existing history of bipolar disorder, psychosis or mental retardation. Cases were aged between 30 and 60 and had two or more episodes of MDD meeting DSM-IV criteria[21] with the first episode occurring between 14 and 50 years of age, and had not abused drugs or alcohol before their first depressive episode. All subjects were interviewed using a computerized assessment system. Interviewers were postgraduate medical students, junior psychiatrists or senior nurses, trained by the CONVERGE team for a minimum of 1 week. The diagnosis of MDD was established with the Composite International Diagnostic Interview (CIDI) (WHO lifetime version 2.1; Chinese version), which used DSM-IV criteria. The interview was originally translated into Mandarin by a team of psychiatrists at Shanghai Mental Health Centre, with the translation reviewed and modified by members of the CONVERGE team.

The replication sample was obtained from five hospitals in the north of China. Patients were diagnosed as having MDD by at least two consultant psychiatrists by DSM-IV criteria. Samples were of both sexes, and all four grandparents were Han Chinese. Cases were aged between 30 and 60, and had two or more episodes of MDD meeting DSM-IV criteria. Exclusion criteria were pregnancy, severe medical conditions, abnormal laboratory baseline values, unstable psychiatric features (for example, suicidal), a history of alcoholism or drug abuse, epilepsy, brain trauma with loss of consciousness, neurological illness, or a concomitant axis I psychiatric disorder. Control subjects were recruited from local communities and provided information about medical and family histories. Exclusion criteria were a history of major psychiatric or neurological disorders, psychiatric treatment or drug abuse, or a family history of severe forms of psychiatric disorders.

The study protocol was approved centrally by the Ethical Review Board of Oxford University (Oxford Tropical Research Ethics Committee) and the ethics committees of all participating hospitals in China. All interviewers were mental health professionals who are well able to judge decisional capacity. The study posed minimal risk (an interview and saliva sample). All participants provided their written informed consent.

DNA sequencing. DNA was extracted from saliva samples using the Oragene protocol. A barcoded library was constructed for each sample. All saliva samples were randomized in allocation to sequencing batches, and experimenters performing the sequencing procedure were blinded to sample allocation and outcome assessment. Sequencing reads obtained from Illumina Hiseq machines were aligned to Genome Reference Consortium Human Build 37 patch release 5 (GRCh37.p5) with Stampy (v1.0.17)[22] using default parameters after filtering out reads containing adaptor sequences or consisting of more than 50% poor quality (base quality ≤ 5) bases. Samtools (v0.1.18)[23] was used to index the alignments in BAM format[23], and Picardtools (v1.62) was used to mark PCR duplicates for downstream filtering. The Genome Analysis Toolkit's (GATK, version 2.6)[24] BaseRecalibrator was then run on the BAM files to create base quality score recalibration tables, masking known SNPs and INDELs from dbSNP (version 137, excluding all sites added after version 129). Base quality recalibration (BQSR) was then performed on the BAM files using GATKlite (v2.2.15)[24] while also removing read pairs that did not have the 'properly aligned segment' bit set by Stampy (1–5% of reads per sample).

Variant calling. Variant discovery and genotyping at all polymorphic SNPs in the 1000G Phase1 East Asian (ASN) reference panel[14] was performed simultaneously using post-BQSR sequencing reads from all samples using the GATK's UnifiedGenotyper (version 2.7-2-g6bda569). We set the option '--genotype_likelihood_model' to 'BOTH', used default annotation outputs for variant calls, and set the '--dbSNP' option in order to use dbSNP v137 rsids to fill in the variant ID column of the output variant call format (VCF) files. Variant quality score recalibration was then performed on these sites using the GATK's VariantRecalibrator (version 2.7-2-g6bda569) and the biallelic SNPs from 1000G Phase1 ASN samples as a true positive set of variants. A sensitivity threshold of 90% to SNPs in the 1000G Phase1 ASN panel was applied for SNP selection for imputation after optimizing for Transition to Transversion (TiTv) ratios in SNPs called. This gave a total of 21,356,798 (9,053,391 known in 1000 Genomes Phase 1 ASN Panel and 11,486,024 novel) biallelic SNPs identified from all chromosomes and unassembled contigs. We put forth 20,539,441 SNPs from the autosomes and chromosome X for imputation of genotype probabilities and downstream analyses.

Genotype likelihood calculation and imputation. Genotype likelihoods (GLs) were calculated at all 20,539,441 SNPs using a sample-specific binomial mixture model implemented in SNPtools (version 1.0)[25], and imputation was performed without a reference panel using BEAGLE (version 3.3.2)[26]. We used BEAGLE to perform imputation, using ten iterations on chunks of 3,000 SNPs with 600 SNPs of overlap. A second round of imputation was performed with BEAGLE on the same GLs, but only at biallelic SNPs polymorphic in the 1000G Phase 1 ASN panel using 572 haplotypes from the 1000 Genomes Phase 1 ASN samples as a reference panel for six iterations on chunks containing roughly 3,000 SNPs with 600 SNPs of overlap. After both rounds of imputation we removed the outer 300 SNPs of every window and ligated imputation results of adjacent chunks. A final set of allele dosages and genotype probabilities was generated from these two sets of imputed results by replacing the results in the former with those in the latter at all sites imputed in the latter. We then applied a conservative set of inclusion thresholds for SNPs for GWAS: (a) P value for violation of the Hardy–Weinberg equilibrium $>10^{-6}$; (b) information score >0.9; (c) MAF in CONVERGE $>0.5\%$, to arrive at the final set of 6,242,619 SNPs for GWAS.

Sample selection for GWAS. Using both processed sequencing data and imputed dosages at SNPs that passed quality control, we assessed the sequencing and imputation quality of all 11,670 samples whose genomic variants we imputed. We first looked into both the nuclear genome and mitochondrial genome for an excess of variants called, since this would indicate cross-sample contamination due to technical issues during sequencing. We quantified the number of singleton variants called in genic regions of the nuclear genome and found a mean of 71.55 private variants per sample that were supported by more than 2 sequencing reads passing sequencing quality controls. We excluded 117 samples with a number of singletons greater than the 99th percentile. Coverage of the mitochondrial genome was, on average, 102×, allowing us to obtain high-quality sequences for this part of the genome. We found a mean of 15.70 heteroplasmic sites per sample, and 116 samples were found to have greater than the 99th percentile of the number of heteroplasmic sites. Of these 116 samples, 26 were already discarded for having excess nuclear genome singletons; and we excluded the remaining 90.

We then checked imputation quality based on the certainty of genotypes imputed (maximum genotype probability >0.9). We identified 29 individuals who had fewer than 90% of their sites with maximum genotype probabilities >0.9. We excluded these samples from further analysis.

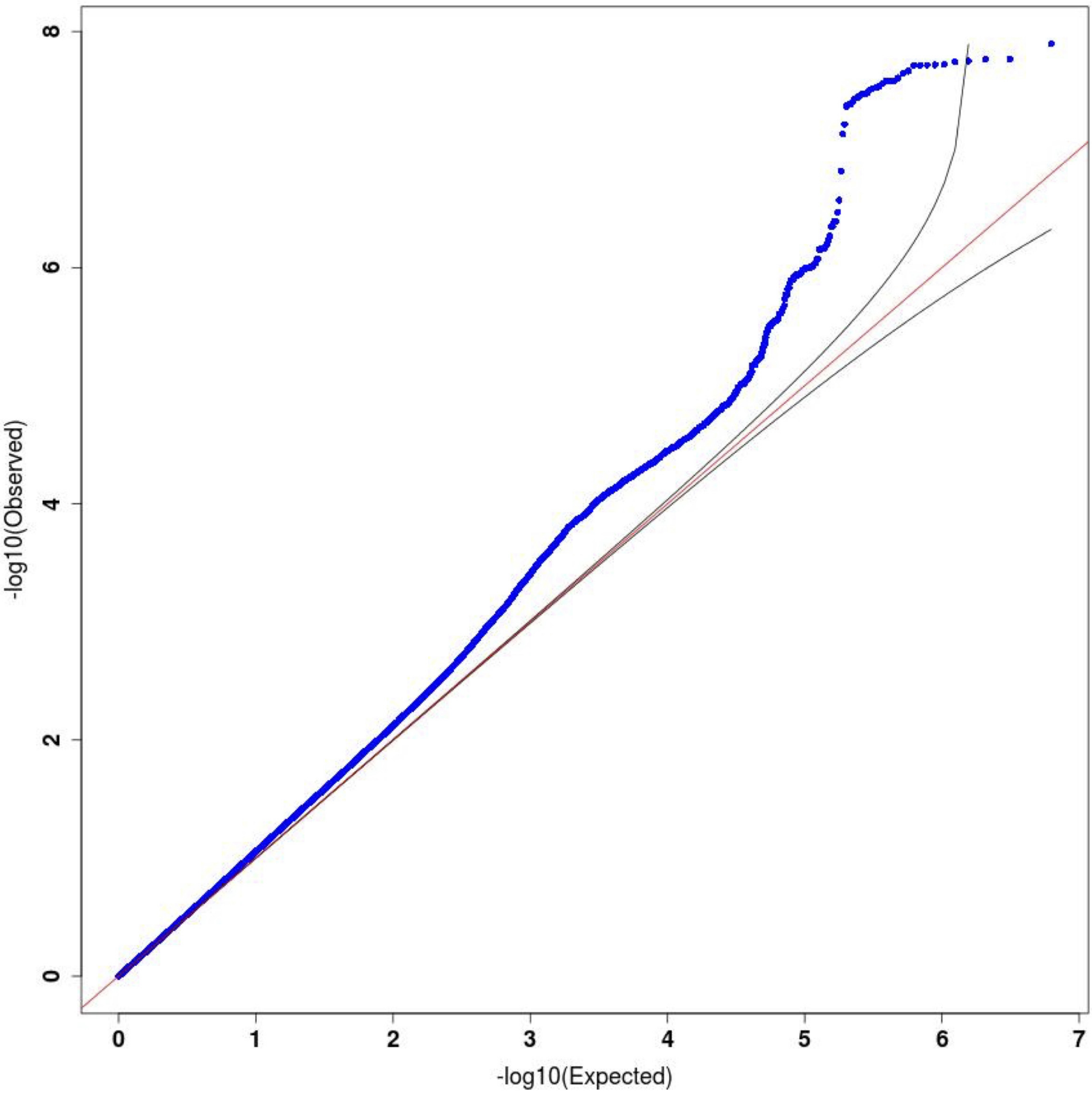
Finally, we assessed the 11,434 remaining samples for genetic relatedness. Although being unrelated to other individuals recruited for the CONVERGE study was a clear criteria in our data collection process, there were instances when the same patient or a relative of the patient visited multiple hospitals and was thus recruited more than once. To exclude duplicates and first-degree relatives from our sample for GWAS, we estimated pairwise genome-wide identity by descent (IBD) using identity by state (IBS) information in hard-called genotypes from imputed genotype probabilities at 399,211 common tagging SNPs across all autosomes ($\text{MAF} > 1\%$, linkage disequilibrium $(\text{LD}) < 0.5$, all known in 1000 Genomes Phase 1). We implemented this in PLINK (v1.07)[27] with the option '--genome'. We excluded a total of 392 samples (duplicates and first-degree relatives) from our final set of samples for GWAS. We retained second-degree relatives and beyond, correcting for the relatedness between them using a linear mixed model. We also excluded 402 samples with incomplete phenotype information, giving a final set of 10,640 samples (5,303 cases of MDD, 5,337 controls) for the primary GWAS of MDD.

GWAS using linear mixed model and liability score estimates. We implemented MLMA using factored spectrally transformed linear mixed models (FastLMM, v2.06.20130802)[9,10] and computed one GRM per chromosome using the mixed linear model with candidate marker excluded (MLMe) approach, removing the SNPs from the chromosome in question from a base set of 322,911 common tagging SNPs from all autosomes ($\text{MAF} > 1\%$, $\text{LD} < 0.5$, all known in 1000G Phase1 ASN panel) to prevent loss of power through 'double fitting' of the candidate SNP (and those in LD with it) in the GRM as a random effect, while testing each SNP as a fixed effect. Manhattan plots and quantile–quantile plots of the $\log_{10}$ of P values of the GWAS were generated with custom code in R (ref. 28). Genomic control inflation factor λ was calculated using custom code in R (ref. 28).

Replication and joint analyses. We genotyped the replication sample on a MassARRAY system mass spectrometer. TYPER4.0 was used to assess the reliability of genotype calls generated by SpectroREAD from the mass spectra. Default genotype call inclusion criteria were used. To perform the association analysis with MDD case–control status at these 12 sites in the replication sample, we obtained effect sizes for discovery from logistic regression with principal component (PC) correction, and then for replication from logistic regression, and then performed fixed-effects meta-analysis.

Polygenic risk profiling and binomial sign-test. Single SNP association results were obtained from the PGC study of MDD[3]. Prior to analysis, SNPs were lifted over to GRCh37/hg19 coordinates and excluded if: (a) monomorphic in either European ($n = 379$) or East Asian ($n = 286$) populations from the 1000 Genomes Project Phase 1 reference data[14]; or (b) absent from the filtered CONVERGE data set. To construct the PGC-trained polygenic score, we initially selected autosomal SNPs with statistical imputation information (information score) greater than 0.9 and MAF greater than 1% in both studies, and performed subsequent LD-based 'clumping' to remove markers from highly correlated SNP pairs (pairwise $r^2 > 0.2$ in East Asians, 500 kb window) while preferentially retaining SNPs with smaller PGC P values. Using the resultant SNP set, we constructed polygene scores based on varying P value thresholds (1×10^{-6}, 1×10^{-5}, 1×10^{-4}, 1×10^{-3}, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, and 1) as previously described[29]. We assessed the predictive value of polygenic scores in a genetically unrelated subset of the CONVERGE sample (with pairwise relatedness less than 0.1) by logistic regression, with adjustment for ancestry principal components, demonstrating significant association with MDD status. The estimated variance in MDD risk accounted for by the polygenic score is given by Nagelkerke's R^2. Using the same P value thresholds, we tabulated the number of independent SNPs with the same direction of allelic effect in the PGC results as observed in CONVERGE. The filtering criteria for SNPs was an information score greater than 0.9 in CONVERGE and MAF greater than 1% in both studies; and an analogous LD-clumping procedure was performed (pairwise $r^2 > 0.2$ in Europeans, 500 kb window). A one-sided binomial sign test was used to assess whether this observed fraction was significantly greater than that expected by chance. Results are given in Extended Data Table 4.

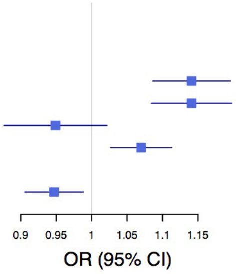
21. Association, A. P. *Diagnostic and statistical manual of mental disorders* 4th edn (American Psychiatric Association, 1994).
22. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21,** 936–939 (2011).
23. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
24. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
25. Wang, Y. *et al.* An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome Res.* **23,** 833–842 (2013).
26. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81,** 1084–1097 (2007).
27. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
28. R Development Core Team. *A language and environment for statistical computing* (R Foundation for Statistical Computing, 2004).
29. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9,** e1003348 (2013).



Extended Data Figure 1 | Quantile–quantile plots for major depressive disorder. Quantile–quantile plot of GWAS for MDD using the mixed linear model with exclusion of the chromosome that the marker is on (MLMe) method implemented in FastLMM on 10,640 samples (5,303 cases, 5,337 controls). Genomic inflation factor $\lambda = 1.070$, rescaled for an equivalent study of 1,000 cases and 1,000 controls (λ_{1000}) = 1.013.

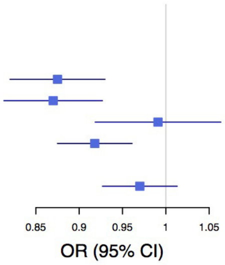
rs2922240 **chr1:11493832:T/C**

study	freq	info	or	se	pval
pri	0.385	1.018	1.141	0.028	2.80E−06
sqnm		1.021	1.141	0.029	5.82E−06
repli	0.386		0.949	0.037	0.15
joint	0.386		1.070	0.022	2.47E−03
pgc	0.495	0.999	0.947	0.021	0.011



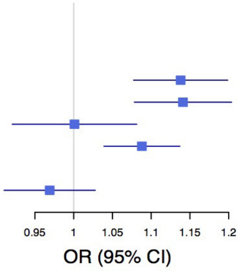
rs3766688 **chr1:175151950:T/C**

study	freq	info	or	se	pval
pri	0.394	1.003	0.875	0.028	1.83E−06
sqnm		1.000	0.870	0.029	1.93E−06
repli	0.388		0.991	0.037	0.81
joint	0.392		0.918	0.022	1.35E−04
pgc	0.546	0.926	0.970	0.022	0.163



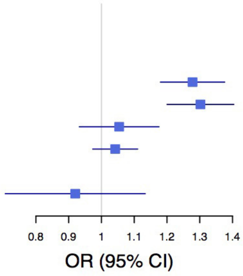
rs57047840 **chr1:228052027:A/G**
*rs10916214(R2:0.479)

study	freq	info	or	se	pval
pri	0.284	0.970	1.138	0.031	4.64E−05
sqnm		0.970	1.141	0.032	4.13E−05
repli	0.302		1.001	0.041	0.99
joint	0.293		1.088	0.025	7.02E−04
pgc*	0.133	0.992	0.969	0.030	0.300



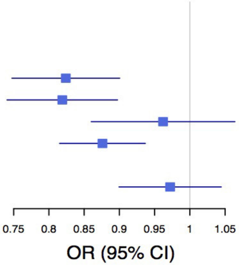
rs55713588 **chr5:9161674:A/G**
*rs13360003(R2:0.177)

study	freq	info	or	se	pval
pri	0.096	0.893	1.278	0.050	6.04E−07
sqnm		0.893	1.302	0.052	3.34E−07
repli	0.100		1.054	0.062	0.39
joint	0.095		1.042	0.035	2.38E−01
pgc*	0.018	0.774	0.920	0.109	0.442



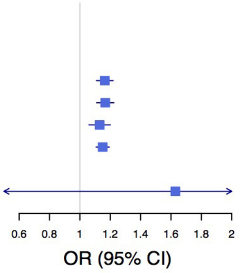
rs55800092 **chr6:4386107:C/T**
*rs17138114(R2:0.891)

study	freq	info	or	se	pval
pri	0.151	1.001	0.824	0.039	1.35E−06
sqnm		1.001	0.819	0.040	6.35E−07
repli	0.138		0.962	0.052	0.45
joint	0.144		0.876	0.031	1.77E−05
pgc*	0.073	0.975	0.972	0.037	0.440



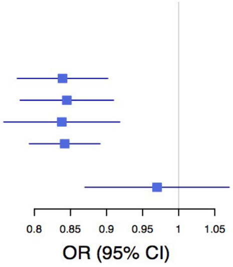
rs12415800 **chr10:69624180:G/A**
*rs16924945(R2:0.323)

study	freq	info	or	se	pval
pri	0.452	0.992	1.164	0.028	1.92E−08
sqnm		0.990	1.167	0.029	8.44E−08
repli	0.443		1.130	0.036	7.71E−04
joint	0.446		1.150	0.022	2.53E−10
pgc*	0.005	0.355	1.629	0.716	0.496



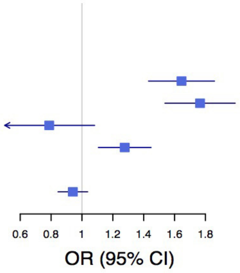
rs35936514 **chr10:126244970:C/T**
*rs35841851(R2:0.981)

study	freq	info	or	se	pval
pri	0.260	0.993	0.839	0.032	1.27E−08
sqnm		0.992	0.845	0.033	2.91E−07
repli	0.261		0.838	0.041	1.68E−05
joint	0.260		0.842	0.025	6.45E−12
pgc*	0.023	0.920	0.970	0.051	0.553



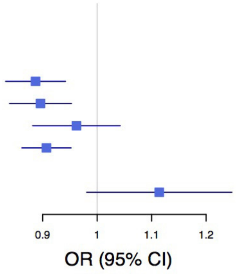
rs61967003 **chr13:107659212:C/T**
*rs16969540(R2:0.980)

study	freq	info	or	se	pval
pri	0.017	0.999	1.645	0.109	6.70E−06
sqnm		0.995	1.765	0.116	9.64E−07
repli	0.015		0.788	0.150	0.11
joint	0.016		1.277	0.087	4.87E−03
pgc*	0.055	0.982	0.941	0.049	0.211



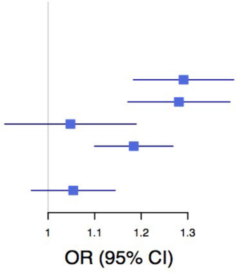
rs17827252 **chr14:66833851:C/G**
*rs2319184(R2:0.965)

study	freq	info	or	se	pval
pri	0.463	1.011	0.887	0.028	1.44E−05
sqnm		1.008	0.896	0.029	1.12E−04
repli	0.458		0.962	0.041	0.34
joint	0.460		0.907	0.023	1.93E−05
pgc*	0.028	0.941	1.114	0.068	0.115



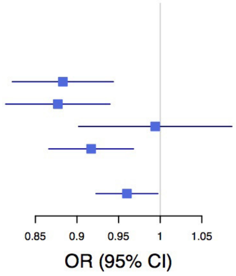
rs11880240 **chr19:34493757:C/G**
*rs7254953(R2:0.645)

study	freq	info	or	se	pval
pri	0.068	1.019	1.291	0.055	8.02E−06
sqnm		1.017	1.281	0.056	1.08E−05
repli	0.070		1.048	0.072	0.51
joint	0.069		1.184	0.043	9.52E−05
pgc*	0.092	0.785	1.054	0.046	0.253



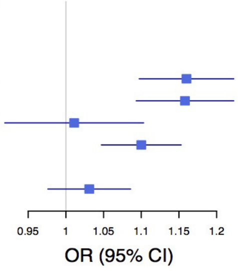
rs1921918 **chrX:24656658:G/A**

study	freq	info	or	se	pval
pri	0.721	0.995	0.883	0.031	3.22E−05
sqnm		0.998	0.877	0.032	3.59E−05
repli	0.733		0.994	0.047	0.9
joint	0.726		0.917	0.026	6.89E−04
pgc	0.581	1.238	0.960	0.019	0.035

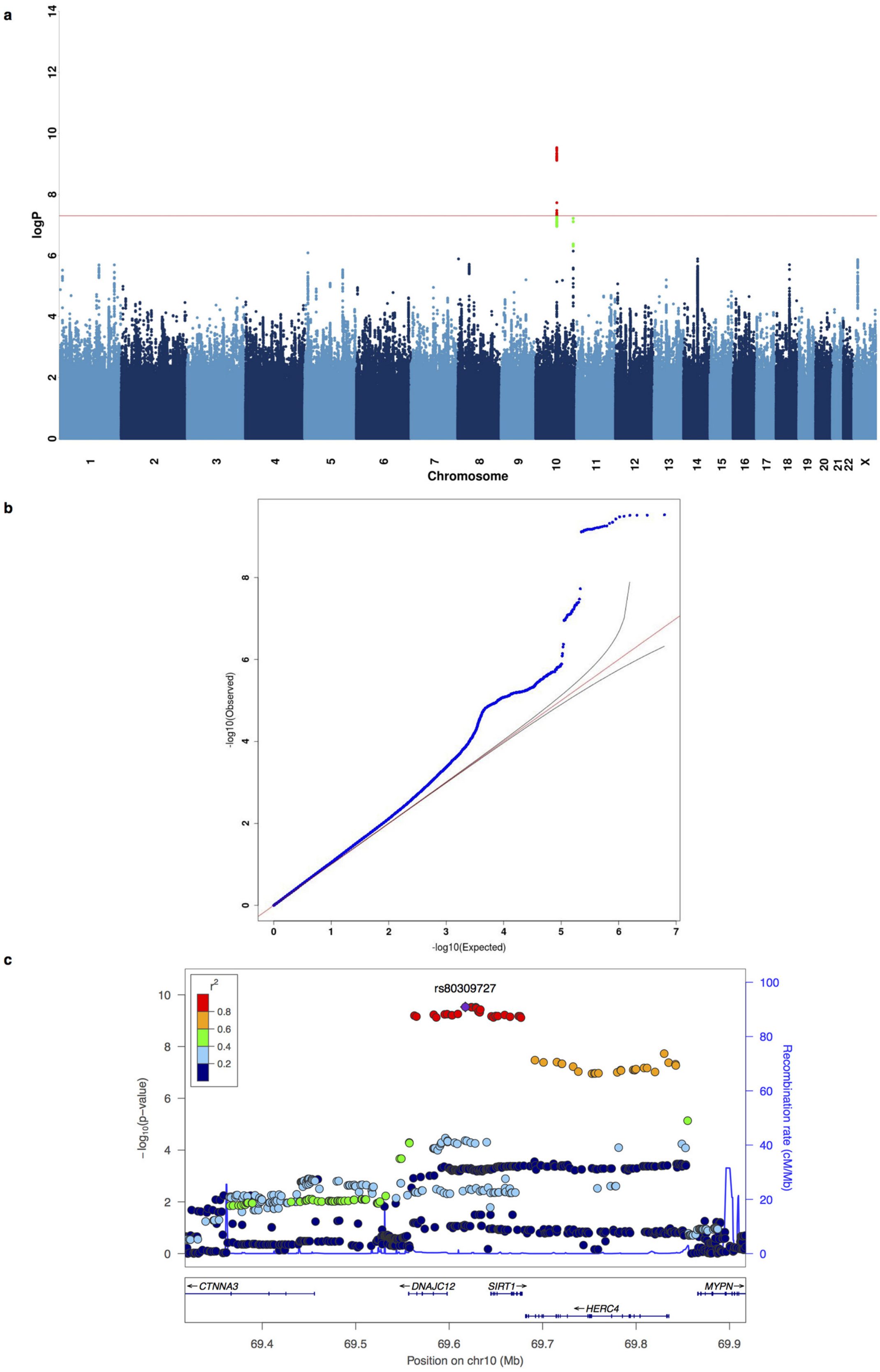


rs11573525 **chrX:25011374:C/T**

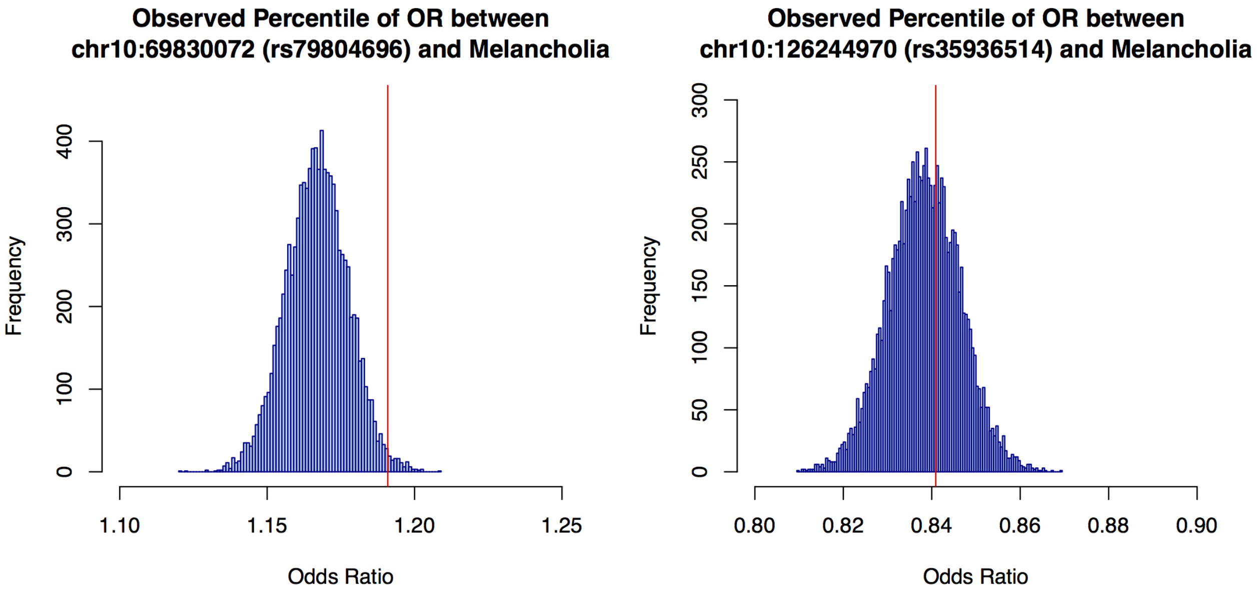
study	freq	info	or	se	pval
pri	0.260	0.971	1.160	0.032	5.86E−06
sqnm		0.975	1.158	0.033	9.60E−06
repli	0.262		1.011	0.047	0.82
joint	0.264		1.100	0.027	3.43E−04
pgc	0.158	1.081	1.031	0.028	0.275



Extended Data Figure 2 | Forest plots of estimated SNP effects in CONVERGE and PGC studies. This figure presents the association odds ratios (OR) at 12 SNPs in CONVERGE and the best available proxy SNPs in PGC-MDD (pairwise $r^2 > 0.6$, 500 kb window; the proxy SNP is marked by an asterisk). We present the alternative allele frequency (freq), odds ratio (or) with respect to the alternative allele, standard error of odds ratio (se) and *P* values of association (pval) for the following analyses (study): primary association analysis with a linear-mixed model using imputed allele dosages in 10,640 samples in CONVERGE (pri); validation analysis with logistic regression model with principal components (PCs) as covariates using genotypes from Sequenom on 9,921 samples in CONVERGE (sqnm); association with MDD with a logistic regression model in a replication cohort of 6,417 samples using genotypes from Sequenom (repli); joint association analysis with MDD with a logistic regression model using imputed allele dosages in CONVERGE and genotypes from Sequenom in a replication cohort (17,057 samples in total; joint).

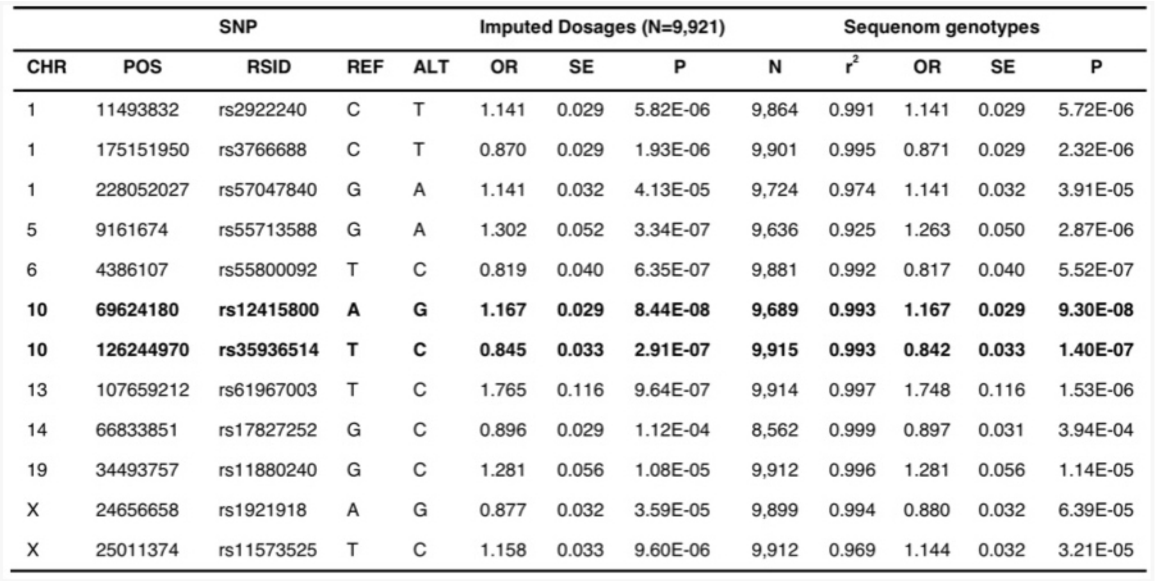


Extended Data Figure 3 | Manhattan and quantile quantile plots for melancholia. **a**, Manhattan plot of GWAS for melancholia using the MLMe method implemented in FastLMM on 9,846 samples (4,509 cases, 5,337 controls). **b**, Quantile–quantile plot of GWAS for melancholia; $\lambda = 1.069$, $\lambda_{1000} = 1.014$. **c**, Regional association plot of GWAS hits on chromosome 10, focusing on top SNP rs80309727 at 5′ of *SIRT1* gene, generated with LocusZoom.



Extended Data Figure 4 | Empirical estimation of the odds ratio increases due to the removal of cases not falling under the diagnostic class of melancholia from an association analysis with major depression. The figures show the empirical distributions of the odds ratios for association with each of two SNPs (rs79804696, rs35936514), after removing a random set of 796 samples, equal to the number of cases of MDD not diagnosed as being melancholic. The horizontal axis is the odds ratio for each analysis, and the vertical axis the frequency of occurrence of the odds ratio in 10,000 analyses. The vertical red line is the observed odds ratio after removing cases of MDD not diagnosed as melancholic.

Extended Data Table 1 | Comparison between association results using imputed dosages and directly genotyped markers



SNP					Imputed Dosages (N=9,921)			Sequenom genotypes				
CHR	POS	RSID	REF	ALT	OR	SE	P	N	r^2	OR	SE	P
1	11493832	rs2922240	C	T	1.141	0.029	5.82E-06	9,864	0.991	1.141	0.029	5.72E-06
1	175151950	rs3766688	C	T	0.870	0.029	1.93E-06	9,901	0.995	0.871	0.029	2.32E-06
1	228052027	rs57047840	G	A	1.141	0.032	4.13E-05	9,724	0.974	1.141	0.032	3.91E-05
5	9161674	rs55713588	G	A	1.302	0.052	3.34E-07	9,636	0.925	1.263	0.050	2.87E-06
6	4386107	rs55800092	T	C	0.819	0.040	6.35E-07	9,881	0.992	0.817	0.040	5.52E-07
10	**69624180**	**rs12415800**	**A**	**G**	**1.167**	**0.029**	**8.44E-08**	**9,689**	**0.993**	**1.167**	**0.029**	**9.30E-08**
10	**126244970**	**rs35936514**	**T**	**C**	**0.845**	**0.033**	**2.91E-07**	**9,915**	**0.993**	**0.842**	**0.033**	**1.40E-07**
13	107659212	rs61967003	T	C	1.765	0.116	9.64E-07	9,914	0.997	1.748	0.116	1.53E-06
14	66833851	rs17827252	G	C	0.896	0.029	1.12E-04	8,562	0.999	0.897	0.031	3.94E-04
19	34493757	rs11880240	G	C	1.281	0.056	1.08E-05	9,912	0.996	1.281	0.056	1.14E-05
X	24656658	rs1921918	A	G	0.877	0.032	3.59E-05	9,899	0.994	0.880	0.032	6.39E-05
X	25011374	rs11573525	T	C	1.158	0.033	9.60E-06	9,912	0.969	1.144	0.032	3.21E-05

The table reports results for association between MDD and 12 SNPs. The first five columns give the chromosome (CHR), genomic position (POS), SNP identifier (RSID), reference allele (REF) on Human Genome Reference GRCh37.p5, and alternative allele (ALT) called in CONVERGE. The next three columns show results for imputed allele dosages at 12 SNPs (odds ratio (OR) of association with MDD with respect to the alternative allele and standard error (SE); *P* values of association (*P*)). The next two columns present the number of samples (N) successfully genotyped using the Sequenom platform (a high-sensitivity and -specificity assay), and the Pearson correlation (r^2) between the imputed allele dosages and the genotypes from Sequenom. The final three columns present results from analyses of association with MDD using genotypes from the Sequenom genotyping platform. Bold type indicates the genome-wide significant markers; Extended Data Table 2 gives further information on the results for these markers.

Extended Data Table 2 | Genotype distribution and *P* values for violation of the Hardy–Weinberg equilibrium in CONVERGE and replication cohorts

MDD Disease State	SNP	CONVERGE		Replication Cohort	
		HomRef/Het/HomAlt	HWE P-value	HomRef/Het/HomAlt	HWE P-value
All	rs12415800	2151/5301/3169	0.445	1212/3037/1920	0.857
	rs35936514	705/4054/5794	0.919	422/2400/3398	0.974
Cases	rs12415800	1178/2626/1490	0.741	654/1538/918	0.829
	rs35936514	318/1919/3027	0.549	190/1136/1783	0.627
Controls	rs12415800	973/2675/1679	0.106	558/1499/1002	0.971
	rs35936514	387/2135/2767	0.389	232/1264/1615	0.503

This table shows the number of samples with the homozygous reference genotype (HomRef), heterozygous genotypes (Het), and homozygous alternative genotype (HomAlt), as well as *P* values for violation of the Hardy–Weinberg equilibrium (HWE) for both CONVERGE study samples and the replication cohort from northern China at the top SNPs rs12415800 in the *SIRT1* locus and rs35936514 in the *LHPP* locus from the GWAS on MDD. The top two rows show these measures for all samples in both the CONVERGE and replication study, the next two rows show these measures for just cases in CONVERGE and the replication cohort, and the last two rows show these measures for just the controls. The genotype distributions for CONVERGE are obtained from hard-called genotypes from maximum imputed genotype probabilities for each sample at each of the two sites. As a genotype will not be called if the maximum genotype probability at a site is lower than 0.9 for any single sample, the total number of CONVERGE samples showing called HomRef/Het/HomAlt genotypes does not equal 10,640 for either SNP. For rs12415800, 19 samples (9 cases, 10 controls) have no genotype calls owing to a maximum genotype probability smaller than 0.9, giving a total of 10,621 CONVERGE (5,294 cases, 5,327 controls) samples with genotype calls. For rs35936514, 87 (39 cases, 48 controls) samples have no genotype calls owing to a maximum genotype probability smaller than 0.9, giving a total of 10,553 (5,264 cases, 5,289 controls) CONVERGE samples with genotype calls.

Extended Data Table 3 | Single-marker association results of top CONVERGE hits in the PGC study of MDD

CONVERGE (10,640 samples)									PGC MDD (18,759 samples)					
CHR	POS	RSID	REF	ALT	FREQ	OR	SE	P	RSID	LD r2	FREQ	INFO	OR	P
1	11493832	rs2922240	C	T	0.3846	1.141	0.028	2.80E-06	rs2922240	1.00	0.495	0.999	0.947	0.011
1	175151950	rs3766688	C	T	0.394	0.875	0.028	1.83E-06	rs3766688	1.00	0.546	0.926	0.970	0.163
1	228052027	rs57047840	G	A	0.2843	1.138	0.031	4.64E-05	rs10916214	0.48	0.133	0.992	0.969	0.300
5	9161674	rs55713588	G	A	0.0956	1.278	0.050	6.04E-07	rs13360003	0.18	0.018	0.774	0.920	0.442
6	4386107	rs55800092	T	C	0.1512	0.824	0.039	1.35E-06	rs17138114	0.89	0.073	0.975	0.972	0.440
10	**69624180**	**rs12415800**	**A**	**G**	**0.4519**	**1.164**	**0.028**	**1.92E-08**	**rs16924945**	**0.32**	**0.005**	**0.355**	**1.629**	**0.496**
10	**126244970**	**rs35936514**	**T**	**C**	**0.2609**	**0.839**	**0.032**	**1.27E-08**	**rs35841851**	**0.98**	**0.023**	**0.92**	**0.970**	**0.553**
13	107659212	rs61967003	T	C	0.0172	1.645	0.109	6.70E-06	rs16969540	0.98	0.055	0.982	0.941	0.211
14	66833851	rs17827252	G	C	0.4624	0.887	0.028	1.44E-05	rs2319184	0.96	0.028	0.941	1.114	0.115
19	34493757	rs11880240	G	C	0.0679	1.291	0.055	8.02E-06	rs7254953	0.65	0.092	0.785	1.054	0.253
X	24656658	rs1921918	A	G	0.7206	0.883	0.031	3.22E-05	rs1921918	1.00	0.581	1.238	0.960	0.035
X	25011374	rs11573525	T	C	0.2602	1.160	0.032	5.86E-06	rs11573525	1.00	0.158	1.081	1.031	0.275

The table compares results from 12 SNPs genotyped in the CONVERGE cohort with either the same SNPs, or best available proxies within a 500 kb window, as reported by the Major Depressive Disorder Working Group of the PGC. The first five columns give the SNP identifier (RSID), chromosome (CHR), genomic position (POS), reference allele (REF) on Human Genome Reference GRCh37.p5, and alternative allele (ALT) called in CONVERGE. The next four columns show the alternative allele frequency (FREQ) and results of association testing with MDD at the 12 SNPs in CONVERGE: odds ratio (OR) of association with MDD with respect to the alternative allele and standard error (SE) in the odds ratio were obtained from a logistic regression model with PCs as covariates; *P* values of association (*P*) were obtained from a linear mixed model with a genetic relatedness matrix containing all samples. The next three columns show the SNP identifier (RSID) of best available proxy of each SNP reported in PGC-MDD, the linkage disequilibrium correlation (LD r^2) expressed as the r^2 value between the SNP in PGC-MDD and SNP in CONVERGE, and the alternative allele frequency (FREQ) at the SNP in PGC-MDD. The last three columns show the information scores (INFO), odds ratios (OR) and *P* values of association with MDD in PGC-MDD from a logistic regression model. Bold type indicates the genome-wide significant markers.

Extended Data Table 4 | Polygenic risk profiling and binomial sign tests

pT	Polygenic risk profiling		Binomial sign test	
	r^2	*P*	No. SNPs (%)	*P*
0.000001	0.000715	**0.0174**	3 (100)	0.125
0.00001	8.40E-05	0.415	12 (66.7)	0.194
0.0001	2.57E-05	0.652	62 (58.1)	0.126
0.001	5.87E-06	0.829	481 (53.6)	0.0605
0.01	8.67E-05	0.407	3632 (51.1)	0.101
0.1	0.00142	0.000797	26106 (50.4)	0.126
0.2	0.00126	0.00156	45166 (50.6)	0.00331
0.3	0.00116	0.00246	61074 (50.5)	0.00627
0.4	0.00125	0.00168	74676 (50.5)	0.00335
0.5	0.0011	0.00317	86429 (50.4)	0.00758
1	0.000924	0.00684	124361 (50.3)	0.0116

The table shows the predictive value of a PGC-trained polygenic risk score on the CONVERGE results. Predictive values are shown at varying *P* value thresholds (*pT*) from $P \leq 1 \times 10^{-6}$ to 1 (that is, all results). *P* is the *P* value of the prediction and r^2 is the amount of variance explained (thus the table shows that including all independent SNPs from the PGC study of MDD, irrespective of individual *P* value, explained 0.09% of MDD risk in CONVERGE.). The number of independent SNPs at each threshold is presented (No. SNPs); the significance of the observed fraction (%) demonstrating a consistent direction of effect was assessed by a one-sided binomial sign test.

doi:10.1038/nature14467

Impermanence of dendritic spines in live adult CA1 hippocampus

Alessio Attardo[1,2]†*, James E. Fitzgerald[1]†* & Mark J. Schnitzer[1,2,3]

The mammalian hippocampus is crucial for episodic memory formation[1] and transiently retains information for about 3–4 weeks in adult mice and longer in humans[2]. Although neuroscientists widely believe that neural synapses are elemental sites of information storage[3], there has been no direct evidence that hippocampal synapses persist for time intervals commensurate with the duration of hippocampal-dependent memory. Here we tested the prediction that the lifetimes of hippocampal synapses match the longevity of hippocampal memory. By using time-lapse two-photon microendoscopy[4] in the CA1 hippocampal area of live mice, we monitored the turnover dynamics of the pyramidal neurons' basal dendritic spines, postsynaptic structures whose turnover dynamics are thought to reflect those of excitatory synaptic connections[5,6]. Strikingly, CA1 spine turnover dynamics differed sharply from those seen previously in the neocortex[7–9]. Mathematical modelling revealed that the data best matched kinetic models with a single population of spines with a mean lifetime of approximately 1–2 weeks. This implies ~100% turnover in ~2–3 times this interval, a near full erasure of the synaptic connectivity pattern. Although *N*-methyl-D-aspartate (NMDA) receptor blockade stabilizes spines in the neocortex[10,11], in CA1 it transiently increased the rate of spine loss and thus lowered spine density. These results reveal that adult neocortical and hippocampal pyramidal neurons have divergent patterns of spine regulation and quantitatively support the idea that the transience of hippocampal-dependent memory directly reflects the turnover dynamics of hippocampal synapses.

The hypothesis that synaptic connectivity patterns encode information has profoundly shaped research on long-term memory. In the hippocampus, synapses in basal CA1 mainly receive inputs from hippocampal area CA3, and the CA3 → CA1 projection has been widely studied regarding its plasticity and key role in memory. As in the neocortex, dendritic spines in the hippocampus are good proxies for excitatory synapses[12], motivating time-lapse imaging of spines as a means of monitoring synaptic turnover[7–10].

Previous work has illustrated *in vivo* imaging of CA1 spines in acute and recently also in chronic preparations[13–15]. We tracked spines for up to ~14 weeks by combining microendoscopes of diffraction-limited resolution[14] (0.85 NA), a chronic mouse preparation for time-lapse imaging in deep brain areas[4], and *Thy1-GFP* mice that express green fluorescent protein (GFP) in a sparse subset of CA1 pyramidal neurons (Fig. 1 and Extended Data Fig. 1). Histological analyses confirmed that this approach induced minimal activation of glia (Extended Data Fig. 2), as shown previously[4,16].

A major concern was that it is not possible to distinguish two or more spines spaced within the resolution limit of two-photon microscopy. This issue is critical for studies of hippocampal spines, which are more densely packed than neocortical spines[17]. To gauge how commonly the appearances of adjacent spines merged together in optical images, we examined tissue slices from *Thy1-GFP* mice, using both two-photon microendoscopy and stimulated-emission depletion (STED) microscopy. STED microscopy offered super-resolution (~70 nm full width at half maximum (FWHM) lateral resolution), an optical resolution nearly nine times finer than that of two-photon microendoscopy[14] (~610 nm), permitting tests comparing pairs of images of the same CA1 dendrites (Fig. 2a and Extended Data Fig. 3).

As expected, we saw nearby spines in STED images that appeared merged in the two-photon images (Fig. 2a). Of 151 spines that appeared unitary in the two-photon images, 23 ± 3.6% (standard error of the mean (s.e.m.)) were actually two spines and 6.0 ± 1.6% were actually three spines ($n = 12$ dendrites) (Fig. 2b). Distances between merged spines in the two-photon images (0.51 ± 0.14 μm; mean ± standard deviation (s.d.)) were below the (0.61 μm) resolution limit[14] (Fig. 2c). Clearly, merging can induce illusory spine stability, since two or more real spines must vanish for a merged spine to disappear.

To treat merging effects quantitatively, we developed a mathematical framework that permits systematic examination of turnover dynamics across different kinetic models and investigation of how the merging of spines on imaging alters the manifestations of these dynamics in two-photon imaging data (Supplementary Information and Extended Data Figs 4–7). We used computer simulations to study how the density and apparent kinetics of merged spines vary with geometric variations of individual spines, spine density, resolution and spine kinetics (Supplementary Information and Extended Data Fig. 8). We also checked experimentally whether fluctuations in spine angle and length, and the radius of the dendrite near the spine, might impact measures of spine turnover (Extended Data Figs 6 and 9). By simulating time-lapse image series, we scored and analysed synthetic data across a broad range of optical conditions, spine densities, geometries and turnover kinetics (Fig. 2d and Extended Data Figs 4 and 5).

The simulations and mathematical modelling confirmed that naive analyses of two-photon data are inappropriate at the spine densities in CA1, owing to merging and the resulting illusion of increased stability (Supplementary Information and Extended Data Figs 4 and 7). For stability analyses, we followed previous studies[7–9] in our use of the survival fraction curve, $S(t)$, the fraction of spines appearing in the initial image acquired at $t = 0$ that also appeared in the image acquired at time t. The shape and asymptotic value of $S(t)$ provide powerful constraints on kinetic models of turnover and the fraction of spines that are permanent (that is, a rate constant of zero for spine loss)[7–9] (Supplementary Information). Strikingly, visual scoring of simulated images (Fig. 2d and Extended Data Fig. 5) yielded underestimates of spine density (Fig. 2e) and patent overestimates of the lifetimes of spines (Fig. 2f). But crucially, our treatment accurately predicted the relationship between the actual density and the visually determined underestimate (Fig. 2e), and properly explained the apparent turnover dynamics, $S(t)$, in terms of the actual kinetics (Fig. 2f and Extended Data Figs 5c, 7). Overall, the simulations showed that face-value interpretations of two-photon images from CA1 are untrustworthy, but

[1]James H. Clark Center for Biomedical Engineering & Sciences, Stanford University, Stanford, California 94305, USA. [2]Howard Hughes Medical Institute, Stanford University, Stanford, California 94305, USA. [3]*CNC* Program, Stanford University, Stanford, California 94305, USA. †Present addresses: Max Planck Institute of Psychiatry, 80804 Munich, Germany (A.A.); Center for Brain Science, Harvard University, Cambridge, Massachusetts 02138, USA (J.E.F.).
*These authors contributed equally to this work.

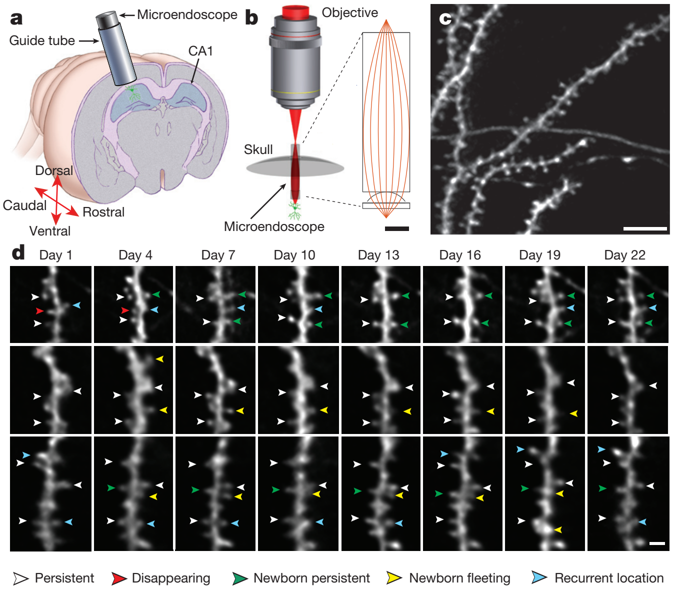


Figure 1 | Dendritic spines are dynamic in CA1 hippocampus of the adult mouse. **a**, A sealed, glass guide tube implanted dorsal to CA1 allows time-lapse *in vivo* imaging of dendritic spines. **b**, A doublet microendoscope projects the laser scanning pattern onto the specimen plane in the tissue. Inset: red lines indicate optical ray trajectories. **c**, CA1 dendritic spines in a live *Thy1-GFP* mouse. **d**, Time-lapse image sequences. Arrowheads indicate spines that either persist across the sequence (white arrowheads), disappear midway (red), arise midway and then persist (green), arise midway and then later disappear (yellow), or disappear and then later appear at an indistinguishable location (cyan). Scale bars: 500 μm (**b**, inset); 10 μm (**c**); 2 μm (**d**).

that it is possible to make quantitatively correct inferences about spine kinetics provided that one properly accounts for the optical resolution. Using the same framework, we next analysed real data.

Initial analyses focused on four mice in which we acquired image stacks of CA1 pyramidal cells and tracked spines every 3 days for 21 days (60 dendrites total; 50 ± 7 (mean ± s.d.) per session) (Fig. 3a). Whenever individual spines appeared at indistinguishable locations on two or more successive sessions, we identified these as observations of the same spine. Overall, we made 4,903 spine observations (613 ± 71 (mean ± s.d.) per day). Spine densities were invariant over time (Fig. 3b) (Wilcoxon signed-rank test; n = 16–50 dendrites per comparison of a pair of days; significance threshold = 0.0018 after Dunn–Šidák correction for 28 comparisons; $P > 0.047$ for all comparisons), as were spine volumes ($P = 0.87$; $n = 43$ spines, Kruskal–Wallis analysis of variance (ANOVA)) (Extended Data Fig. 10) and the turnover ratio, the fraction of spines arising or vanishing since the last session[7] (Fig. 3b) (Wilcoxon signed-rank test; 14–40 dendrites; significance threshold = 0.0025 after correction for 20 comparisons; $P > 0.31$ for all comparisons). Fewer than 50% of spines (46 ± 2%; mean ± s.e.m.; n = 4 mice) were seen throughout the experiment (Extended Data Fig. 1d, e), although our simulations had shown that this naive observation overestimated spine stability.

The time invariance of spine densities and turnover ratios implied that through our mathematical framework we could determine the underlying kinetic parameters governing turnover. $S(t)$ curves for the total and newborn spine populations ostensibly resembled those reported for the neocortex[7–9] (Fig. 3c). However, unlike in the neocortex, the 46% of spines seen in all sessions differed notably from the odds that a spine was observed twice in the same location across two distant time points, as quantified by the asymptotic value of $S(t)$ (73 ± 3%). This discrepancy suggested that, as our modelling had indicated, many CA1 spines might vanish and reappear in an ongoing way at indistinguishable locations.

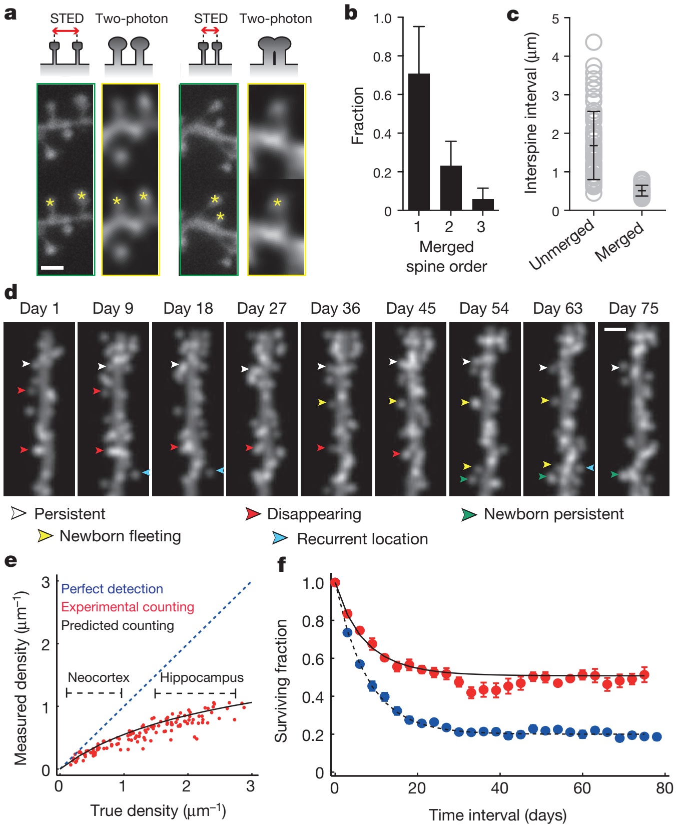


Figure 2 | A simple kinetic model is sufficient to describe CA1 pyramidal cell spine dynamics. **a**, Two-photon microendoscopy and STED imaging of the same dendrites *in vitro*. Top, two-photon images depict spines closer than the resolution limit as merged entities. Bottom, asterisks mark example visually scored spines, showing cases in which nearby spines do (right) or do not (left) merge. **b**, Fraction of spines (n = 151 total) seen by two-photon imaging that were one, two or three spines as determined by STED imaging. Black bars show mean ± s.d. for 12 dendrites. **c**, Separations between adjacent unmerged spines and pairs of spines that appeared merged by two-photon imaging. Open grey circles mark individual results from each of n = 150 spines. Black bars show mean ± s.d. **d**, Example computer-simulated, time-lapse image sequence used to quantify how resolution limits impact measured spine densities and dynamics. **e**, Computational modelling predicts the underestimation of spine density due to the finite optical resolution. Blue diagonal line: perfect detection of all spines. Black horizontal dashed lines: typical ranges of spine densities on pyramidal cells in neocortex and hippocampus. Red data: results from visually scoring simulated images of dendrites of varying spine densities. Black curve: prediction from the scoring model using 600 nm as the minimum separation between two spines correctly distinguished. **f**, Modelling predicts the overestimation of spine stability due to merging of adjacent spines in resolution-limited images. Blue data: survival fraction values (mean ± s.e.m.) for actual spine turnover in computer simulations (spine density: 2.56 μm^{-1}). Red data: apparent turnover for these same simulated dendrites, as scored from simulated two-photon images. Black curves: theoretical predictions for spine survival based on the scoring model. Scale bars: 1 μm (**a**); and 2 μm (**d**).

We next tested whether a prolonged environmental enrichment would alter spine turnover. Previous data from rats have indicated that basal CA1 spine density can rise ~10% after environmental enrichment[18]. Data from mice are limited to CA1 apical dendrites and have yielded contradictory results[19,20] (Supplementary Discussion). In three mice we imaged a total of 55 basal dendrites (39 ± 14 (s.d.) per day) across 16 sessions, 3 days apart (Fig. 3d). After session 8 we moved the mice to an enriched environment, where they stayed throughout sessions 9–16. We made 8,727 spine observations in total (545 ± 216 (s.d.) per day).

Comparisons of baseline and enriched conditions revealed no differences in spine density or turnover (Fig. 3e) (n = 10–53 dendrites;

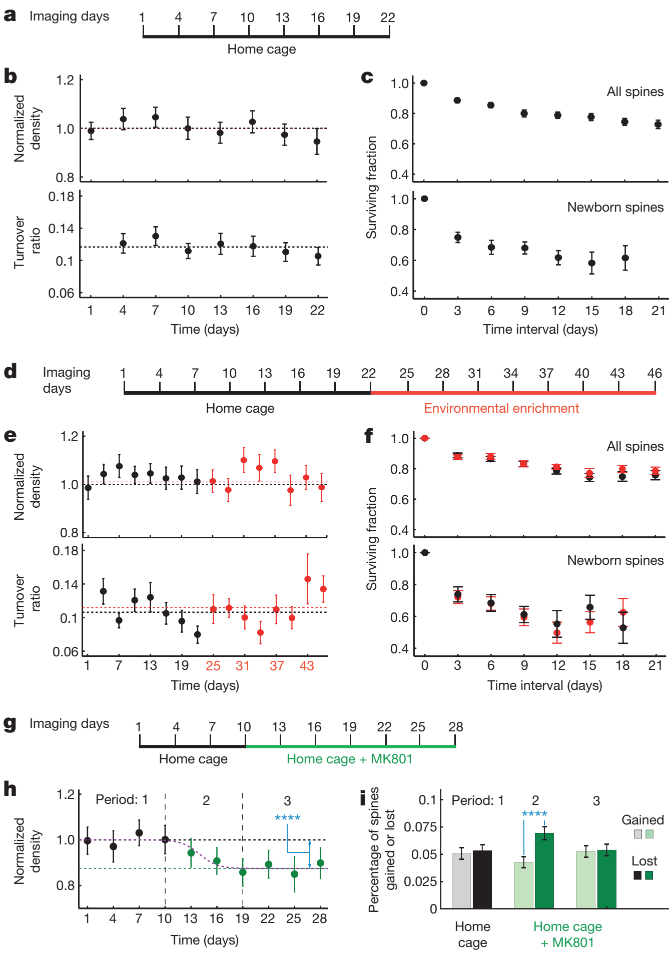


Figure 3 | NMDA receptor blockade, but not environmental enrichment, altered spine turnover dynamics. **a**, Schedule of baseline imaging sessions. **b**, Neither measured spine densities (top) nor turnover ratios (bottom) varied for mice in their home cages. Horizontal lines: mean spine density and turnover ratio. **c**, Spine survival (top) and newborn spine survival (bottom). **d**, Schedule for study on environmental enrichment. **e**, **f**, No significant differences existed between baseline (black points) and enriched conditions (red) regarding spine density (**e**, top), turnover (**e**, bottom), survival (**f**, top), or newborn spine survival, (**f**, bottom). **g**, Schedule for the study on NMDA receptor blockade. **h**, MK801 caused a significant decline in spine density. Data are from mice imaged four times before (black) and six times during (green) MK801 administration. Black and green horizontal lines respectively indicate mean densities during baseline and on the last 4 days of MK801 treatment. The density decrease was highly significant (Wilcoxon signed-rank test; $n = 29$ dendrites; $P = 0.0007$). **i**, The decline in spine density early in MK801 treatment arose from a transient, highly significant difference between the rates at which spines were lost (darker bars) and gained (lighter bars) (Wilcoxon signed-rank test; $n = 29$ dendrites; $P = 0.0008$). Greyscale and green-shaded bars represent percentages of spines gained and lost for sessions before and during MK801 dosage. ****$P < 0.001$. We normalized spine densities to their mean values in baseline conditions, which were 1.03 μm^{-1} (**b**, top), 0.90 μm^{-1} (**e**, top) and 0.78 μm^{-1} (**h**). All error bars are s.e.m. for dendrites.

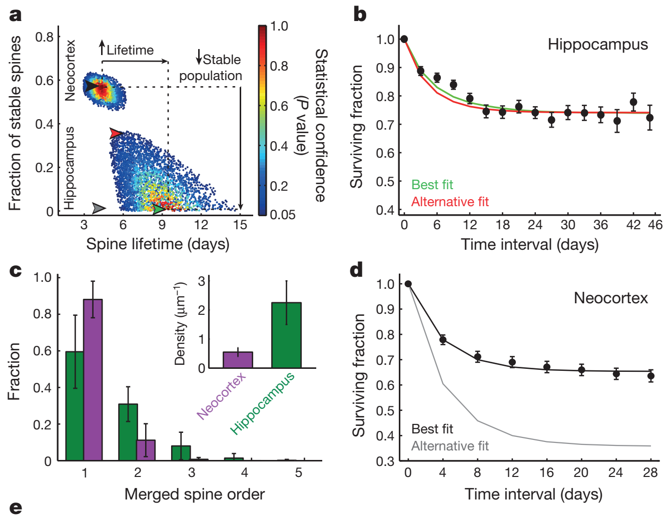
$P > 0.10$, 8 paired comparisons of density; $P > 0.039$, 7 paired comparisons of turnover ratio; Wilcoxon signed-rank tests with significance thresholds of 0.006 and 0.007, respectively, after corrections for multiple comparisons). Neither were there differences in spine survival ($P > 0.057$; 7 time points; $n = 10$–49 dendrites; Wilcoxon signed-rank test; significance threshold of 0.007 after Dunn–Šidák correction), nor in newborn spine survival ($P > 0.29$; 6 time points; $n = 5$–35 dendrites; significance threshold of 0.008; Fig. 3f). Thus, in mice, continuous enrichment does not substantially alter spine dynamics on CA1 basal dendrites. Nevertheless, mean volumes of stable spines underwent a slight (7 ± 3% (s.e.m.)) but significant decline upon enrichment (Wilcoxon signed-rank test; $P = 0.007$; 60 spines tracked for 16 sessions) (Extended Data Fig. 10d). Data on the structural effects of long-term potentiation (LTP) suggest an explanation of these findings; CA1 spine densities rise transiently after LTP induction but return to baseline values 2 h later[21], implying that continual enrichment would cause no net change in spine densities (Supplementary Discussion).

We next examined whether blockade of NMDA glutamate receptors impacts spine turnover. These receptors are involved in multiple forms of neural plasticity, including in the CA1 area[22]. In the neocortex, NMDA receptor blockade stabilizes spines by slowing their elimination while keeping their formation rate unchanged[10]. We tracked CA1 spines across 10 sessions at 3-day intervals in mice receiving the NMDA receptor blocker MK801 beginning after session 4 and onward (Fig. 3g). We examined 26 dendrites (25 ± 1.4 (s.d.) per session), made 5,020 spine observations (502 ± 32 (s.d.) per day), and found that MK801 induced a significant decline (12 ± 3% (s.e.m.)) in spine density (Wilcoxon signed-rank test; 25 dendrites; $P = 0.0007$) (Fig. 3h). This stemmed from a transient disparity in the rates of spine loss versus gain (loss rate was 215 ± 42% (s.e.m.) of the rate of gain; Wilcoxon signed-rank test; 25 dendrites; $P = 0.0008$) (Fig. 3i). These results indicate that the survival odds of CA1 spines depend on NMDA receptor function, illustrate our ability to detect changes in spine dynamics, and show that CA1 and neocortical spines have divergent responses to NMDA receptor blockade.

To ascertain the underlying time constants governing spine turnover, we compared the $S(t)$ curves of visually scored spines to predictions from a wide range of candidate kinetic models (Fig. 4a). In each model there was a subset (0–100%) of permanent spines; the remaining spines were impermanent, with a characteristic lifetime, τ. Since environmental enrichment left the observed spine dynamics unchanged, we pooled the baseline and enriched data sets to extend the analyses to longer time-scales (Fig. 4b). By varying the actual spine density, fraction of stable spines, and characteristic lifetime for the unstable fraction, we identified the model that best fit the $S(t)$ curves, using a maximum likelihood criterion (Supplementary Information). This model had 100% impermanent spines, with an actual density of 2.6 μm^{-1} and τ of ~10 days (Fig. 4a, b). This is twice the ~5-day lifetime reported for the transient subset of neocortical spines[7–9]. There were also models with both permanent and impermanent spines that gave reasonable, albeit poorer fits to the CA1 data (Fig. 4a, b). Crucially, our analysis identified all models whose fits were significantly worse than the best model (white regions in Fig. 4a; $P < 0.05$, likelihood-ratio test); we regarded these as unsatisfactory in accounting for the CA1 data (Supplementary Information).

Our results pointing to a single population of unstable spines in CA1 contrast markedly with findings in adult neocortex, where >50% of spines seem permanent[7–9]. To make even-handed comparisons, we used our framework to re-analyse published data acquired in the mouse somatosensory neocortex that had supported this conclusion[7]. Owing to the lower density of neocortical spines, merging is far less of a concern (Fig. 4c), and our modelling confirmed that ~60% of neocortical spines are stable over very long timescales, supporting past conclusions[7].

Nevertheless, we found very significant differences between CA1 and neocortical spine turnover dynamics (Fig. 4a; $P = 0.01$, likelihood ratio test). Models with only impermanent spines, which well described CA1, were insufficient ($P < 10^{-62}$, likelihood ratio test) for neocortex (Fig. 4a, d). The discrepant lifetimes of impermanent spines in the two areas (~10 days (CA1) versus ~5 days (neocortex)) posed further incompatibility (Fig. 4a, e). Conversely, models that explained neocortical spine turnover were incompatible with the CA1 data ($P = 0.01$, likelihood ratio test). Modelling alone cannot eliminate the possibility that CA1 basal dendrites have some permanent spines, but if any such spines exist they compose a far smaller fraction than in the neocortex. Hence, CA1 and the neocortex have distinct spine dynamics, percentages of impermanent spines, and turnover time constants.



e

	Hippocampus (basal CA1)	Neocortex (S1 and M1)	
Populations	100%	60%	40%
Lifetime	Medium (5–15 days)	Long (entire life)	Short (~5 days)
NMDAr block	↑ Turnover dynamics	↓ Turnover dynamics	
Experience/ learning	No detected change in dynamics	↓ Pre-existing spine lifetime	↑ Newborn spine lifetime

Figure 4 | CA1 and neocortical spines exhibit distinct turnover kinetics. **a**, Multiple kinetic models are consistent with data on spine survival. Each model considered had two subpopulations: permanent and impermanent spines. Abscissa: mean lifetime for impermanent spines. Ordinate: fraction of spines that are permanent. Each datum is for an individual model; colour denotes the level of statistical significance at which the model could be rejected. Red points denote models that best fit the data. No results are shown for models incompatible with the data ($P < 0.05$). Models that best fit data from CA1 have ~100% impermanent spines, with a ~10 day lifetime (green arrowhead). There are also models with permanent subpopulations that cannot be statistically rejected (for example, red arrowhead). Models that best fit patterns of neocortical spine turnover (black arrowhead), from mice age- and gender-matched[7] to those used here, have ~50–60% permanent spines and a shorter lifetime (~5 days) for impermanent spines than in CA1. Models lacking permanent spines poorly fit the neocortical data; grey arrowhead marks the model for the grey curve in **d**. The four arrowheads indicate the models that generated the colour-corresponding curve fits in **b**, **d**. **b**, Empirically determined survival curve for CA1 spines (black data: mean ± s.e.m.; data set of Fig. 3e, f) over 46 days, compared to predictions (solid curves) from two of the models in **a** (green and red arrowheads in **a**). Green curve: best-fitting model, which has no permanent spines. Red curve: an example model with both stable and unstable spines. **c**, Owing to the higher density of spines in CA1 than in the neocortex (inset), optical merging is far more common in CA1. Given what appears to be one spine, vertical bars represent the probability as determined from the computational model that the observation is actually of 1–5 spines. Probabilities were calculated using spine density values of the inset, 0.75 μm as the minimum separation, L, needed to distinguish adjacent spines. Error bars: range of results for L within 0.5–1.0 μm. **d**, Empirically determined survival curves for neocortical spines (data set from ref. 7) over 28 days, compared with predictions (solid curves) from two different models for spine turnover in **a** (grey and black arrowheads in **a**). Black curve: best fit attained with a stable subpopulation of spines. Grey curve: a model lacking permanent spines, poorly fitting the data. **e**, Spines in CA1 and neocortex differ substantially in proportions of permanent versus impermanent spines, spine lifetimes, effects of NMDA receptor blockade, and learning or novel experience.

Further distinguishing CA1 and the neocortex are the contrary roles of NMDA receptor blockade (Figs 3h, i and 4e). In the neocortex, MK801 promotes stability by decreasing spine loss[10] and blocking spine addition[11]; conversely, NMDA receptor activation may speed turnover via addition of new spines and removal of pre-existing neocortical spines supporting older memories. In CA1, MK801 speeds turnover and promotes instability (Fig. 3h, i), suggesting that NMDA receptor activation may transiently slow turnover and stabilize spines. Indeed, LTP induction in CA1 is associated with stabilization of existing spines and growth of new spines[5,23,24].

A natural interpretation is that spine dynamics may be specialized by brain area to suit the duration of information retention. The neocortex, a more permanent repository, might need long-lasting spines for permanent information storage and shorter-lasting ones ready to be stabilized if needed[8,9]. The hippocampus, an apparently transient repository of information, might only require transient spines. The ~1–2-week mean lifetime for the ~100% impermanent CA1 spines implies a near full erasure of synaptic connectivity patterns in ~3–6 weeks, matching the durations that spatial and episodic memories are hippocampal-dependent in rodents[1,2] (but see also ref. 25). The ensemble place codes of CA1 neurons also refresh in ~1 month[16], which could arise from turnover of the cells' synaptic inputs. Since 75–80% of CA3 → CA1 inputs are monosynaptic[26], CA1 spine impermanence probably implies a continuous re-patterning of CA3 → CA1 connectivity throughout adulthood, which, owing to the sheer number of synapses and sparse connections is unlikely to assume the same configuration twice.

Supporting these interpretations, artificial neural networks often show a correspondence between synapse lifetime and memory longevity[27], although in some models spine turnover and memory erasure can be dissociated[28,29]. Computational studies also show that elimination of old synapses can enhance memory capacity[29,30]. More broadly, networks that can alter synaptic lifetimes, not just connection strengths, can more stably store long-term memories while rapidly encoding new ones[27].

The data described here are consistent with a single class of CA1 spines, but future studies should examine both the finer kinetic features and cellular or network mechanisms of turnover[23]. By using fluorescence tags to mark spines undergoing plastic changes, *in vivo* imaging might help relate connection strengths, spine lifetimes and memory performance. Finally, researchers should investigate spine turnover in animals as they learn to perform a hippocampal-dependent behaviour, to build on the results here by looking for direct relationships between CA1 spine stability and learning.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 August 2014; accepted 10 April 2015.
Published online 22 June 2015.

1. Squire, L. R. & Zola-Morgan, S. The medial temporal lobe memory system. *Science* **253,** 1380–1386 (1991).
2. Frankland, P. W. & Bontempi, B. The organization of recent and remote memories. *Nature Rev. Neurosci.* **6,** 119–130 (2005).
3. Frey, U. & Morris, R. G. Synaptic tagging and long-term potentiation. *Nature* **385,** 533–536 (1997).
4. Barretto, R. P. *et al.* Time-lapse imaging of disease progression in deep brain areas using fluorescence microendoscopy. *Nature Med.* **17,** 223–228 (2011).
5. Engert, F. & Bonhoeffer, T. Dendritic spine changes associated with hippocampal long-term synaptic plasticity. *Nature* **399,** 66–70 (1999).
6. Maletic-Savatic, M., Malinow, R. & Svoboda, K. Rapid dendritic morphogenesis in CA1 hippocampal dendrites induced by synaptic activity. *Science* **283,** 1923–1927 (1999).
7. Holtmaat, A. J. *et al.* Transient and persistent dendritic spines in the neocortex *in vivo. Neuron* **45,** 279–291 (2005).
8. Xu, T. *et al.* Rapid formation and selective stabilization of synapses for enduring motor memories. *Nature* **462,** 915–919 (2009).
9. Yang, G., Pan, F. & Gan, W. B. Stably maintained dendritic spines are associated with lifelong memories. *Nature* **462,** 920–924 (2009).
10. Zuo, Y., Yang, G., Kwon, E. & Gan, W. B. Long-term sensory deprivation prevents dendritic spine loss in primary somatosensory cortex. *Nature* **436,** 261–265 (2005).
11. Yang, G. *et al.* Sleep promotes branch-specific formation of dendritic spines after learning. *Science* **344,** 1173–1178 (2014).
12. Harris, K. M. Structure, development, and plasticity of dendritic spines. *Curr. Opin. Neurobiol.* **9,** 343–348 (1999).
13. Mizrahi, A., Crowley, J. C., Shtoyerman, E. & Katz, L. C. High-resolution *in vivo* imaging of hippocampal dendrites and spines. *J. Neurosci.* **24,** 3147–3151 (2004).

14. Barretto, R. P., Messerschmidt, B. & Schnitzer, M. J. *In vivo* fluorescence imaging with high-resolution microlenses. *Nature Methods* **6,** 511–512 (2009).
15. Gu, L. *et al.* Long-term *in vivo* imaging of dendritic spines in the hippocampus reveals structural plasticity. *J. Neurosci.* **34,** 13948–13953 (2014).
16. Ziv, Y. *et al.* Long-term dynamics of CA1 hippocampal place codes. *Nature Neurosci.* **16,** 264–266 (2013).
17. Harris, K. M. & Stevens, J. K. Dendritic spines of CA 1 pyramidal cells in the rat hippocampus: serial electron microscopy with reference to their biophysical characteristics. *J. Neurosci.* **9,** 2982–2997 (1989).
18. Moser, M. B., Trommald, M. & Andersen, P. An increase in dendritic spine density on hippocampal CA1 pyramidal cells following spatial learning in adult rats suggests the formation of new synapses. *Proc. Natl Acad. Sci. USA* **91,** 12673–12675 (1994).
19. Rampon, C. *et al.* Enrichment induces structural changes and recovery from nonspatial memory deficits in CA1 NMDAR1-knockout mice. *Nature Neurosci.* **3,** 238–244 (2000).
20. Sanders, J., Cowansage, K., Baumgartel, K. & Mayford, M. Elimination of dendritic spines with long-term memory is specific to active circuits. *J. Neurosci.* **32,** 12570–12578 (2012).
21. Bourne, J. N. & Harris, K. M. Coordination of size and number of excitatory and inhibitory synapses results in a balanced structural plasticity along mature hippocampal CA1 dendrites during LTP. *Hippocampus* **21,** 354–373 (2011).
22. Huerta, P. T., Sun, L. D., Wilson, M. A. & Tonegawa, S. Formation of temporal memory requires NMDA receptors within CA1 pyramidal neurons. *Neuron* **25,** 473–480 (2000).
23. Yasumatsu, N., Matsuzaki, M., Miyazaki, T., Noguchi, J. & Kasai, H. Principles of long-term dynamics of dendritic spines. *J. Neurosci.* **28,** 13592–13608 (2008).
24. Toni, N., Buchs, P. A., Nikonenko, I., Bron, C. R. & Muller, D. LTP promotes formation of multiple spine synapses between a single axon terminal and a dendrite. *Nature* **402,** 421–425 (1999).
25. Goshen, I. *et al.* Dynamics of retrieval strategies for remote memories. *Cell* **147,** 678–689 (2011).
26. Sorra, K. E. & Harris, K. M. Stability in synapse number and size at 2 hr after long-term potentiation in hippocampal area CA1. *J. Neurosci.* **18,** 658–671 (1998).
27. Fusi, S., Drew, P. J. & Abbott, L. F. Cascade models of synaptically stored memories. *Neuron* **45,** 599–611 (2005).
28. Abraham, W. C. & Robins, A. Memory retention—the synaptic stability versus plasticity dilemma. *Trends Neurosci.* **28,** 73–78 (2005).
29. Wu, X. E. & Mel, B. W. Capacity-enhancing synaptic learning rules in a medial temporal lobe online learning model. *Neuron* **62,** 31–41 (2009).
30. Poirazi, P. & Mel, B. W. Impact of active dendrites and structural plasticity on the memory capacity of neural tissue. *Neuron* **29,** 779–796 (2001).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Li, J. Lecoq and E. T. W. Ho for technical assistance, R. P. J. Barretto and B. Messerschmidt for advice, and A. Holtmaat for providing published data sets. The National Institute of Mental Health, The National Institute on Aging, and the Ellison Foundation provided grants (M.J.S.). Fellowships from the National Science Foundation and The Stanford Center for Mind, Brain, and Computation provided graduate research support (J.E.F.).

Author Contributions A.A. and M.J.S. designed experiments. A.A. performed experiments. A.A. and J.E.F. analysed data. J.E.F. designed and performed the modelling. A.A., J.E.F. and M.J.S. wrote the paper. M.J.S. supervised the research.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.J.S. (mschnitz@stanford.edu).

METHODS

Animals and surgical preparation. Stanford University's Administrative Panel on Laboratory Animal Care approved all procedures. We imaged neurons in mice expressing GFP driven by the *Thy1* promoter[31] (heterozygous males 10–12 weeks old, GFP-M lines on a C57BL/6 × F1 background). We did not perform any formal randomization in the assignment of mice to specific groups, but we informally selected mice in a random manner without use of any exclusion criteria. We performed surgeries as previously published[4] but with a few modifications. We anaesthetized mice using isoflurane (1.5–3% in O_2) and implanted a stainless steel screw into the cranium above the brain's right hemisphere. We performed a craniotomy in the left hemisphere (2.0 ± 0.3 mm posterior to bregma, 2.0 ± 0.3 mm lateral to midline) using a 1.8-mm-diameter trephine and implanted the optical guide tube with its window just dorsal to, but not within, area CA1, preserving the alveus.

Guide tubes and microlenses. Guide tubes were glass capillaries (1.5 mm (ID), 1.8 mm (OD) or 2 mm (ID), 2.4 mm (OD); 2–3 mm in length). We attached a circular coverslip, matched in diameter to that of the capillary's outer edge, to one end of the guide tube by using optical epoxy (Norland Optical Adhesive 81). We used 1.0-mm-diameter micro-optical probes of diffraction-limited resolution (0.8 NA, 250 μm working distance in water) that were encased in a 1.4-mm-diameter sheath[14].

***In vivo* two-photon imaging.** We used a modified commercial two-photon microscope (Prairie Technologies) equipped with a tuneable Ti:Sapphire laser (Chameleon, Coherent). We tuned the laser emission to 920 nm and adjusted the average illumination power at the sample (~5–25 mW) for consistency in signal strength across imaging sessions in each mouse. For microendoscopy we used a 20 × 0.8 NA objective (Zeiss, Plan-Apochromat) to deliver illumination into the microlenses. In some cases we imaged directly through the glass cannula using a Olympus LUMPlan Fl/IR 0.8 NA ×40 water immersion objective lens and confirmed that the optical resolution in all three spatial axes was identical between the two approaches. Beginning at 15–18 days after surgery, we imaged mice every 3 days under isoflurane anaesthesia (1.5% in O_2) for a total of 8–16 sessions each lasting 60–90 min. We imaged some mice at irregular intervals up to 80 days.

MK801 treatment. In mice subject to the protocol of Fig. 2d, after the fourth imaging session we administered MK801 (Tocris Bioscience; 0.25 mg g^{-1} body weight; dissolved in saline) in two intraperitoneal injections each day (8–10 h apart) as described previously[10].

Enriched environment. Animals given an enriched environment had a larger cage (42 (length) × 21.5 (width) × 21.5 (height) cm^3) that contained a running wheel, objects of various colours, textures and shapes, plastic tunnels, and food with different flavours. We changed the objects, as well as their placements within the cage, every 3–4 days to encourage exploration and maintain novelty. We provided food and water *ad libitum*.

Histology. At the end of *in vivo* experimentation, we deeply anaesthetized mice with ketamine (100 mg kg^{-1}) and xylazine (20 mg kg^{-1}). We then perfused PBS (pH 7.4) into the heart, followed by 4% paraformaldehyde in PBS. We fixed brains overnight at 4 °C and prepared floating sections (50 mm) on a vibrating microtome (VT1000S, Leica). Before *in vitro* imaging of GFP fluorescence, we washed the fluorescent sections with PBS buffer several times and quenched them by incubation in 150 mM glycine in PBS for 15 min. After three washes in PBS, we mounted sections with Fluoromout-G (Southern Biotech). We inspected the sections using either two-photon fluorescence imaging or a STED microscope (Leica TCS STED CW, equipped with a Leica HCX PL APO 100 × 1.40 NA oil-immersion objective.).

For immunostaining sections were washed with PBS buffer several times before quenching and permeabilization (15 min incubation in 0.1% Triton-X in PBS). Sections were incubated in blocking solution (1% Triton X-100, 2% BSA, 2% goat serum in PBS) for 4 h. Primary antibodies (rat anti-CD68, FA11-ab5344, Abcam, 1:100 dilution; mouse anti-GFAP, MAB3402, Millipore, 1:500 dilution) were diluted in blocking solution and sections were incubated overnight in this solution. The following day sections were washed with PBS and incubated in diluted secondary antibody: Cy5-conjugated goat anti-mouse IgG, A10524 and Cy5-conjugated goat anti-rat IgG, A10525; both Molecular Probes; both 1:1,000 dilution, in blocking solution for 3 h. All staining procedures were done at room temperature. After three washes in PBS, sections were mounted with Fluoromout-G (Southern Biotech). Histological specimens were inspected on a confocal fluorescence microscope (Leica SP2 AOBS).

Imaging sessions. We mounted the isoflurane-anaesthetized mice on a stereotactic frame. To perform microendoscopy we fully inserted the microendoscope probe into the guide tube such that the probe rested on the guide tube's glass window. To attain precise and reliable three-dimensional alignments across all imaging sessions of the brain tissue undergoing imaging, we used a laser-based alignment method. We positioned a laser beam such that when the mouse's head was properly aligned, the beam reflected off the back surface of the microendoscope and hit a designated target. This ensured that at each imaging session the long axis of the microendoscope was perpendicular to the optical table to within ~1 angular degree. Otherwise we inserted a drop of water in the cannula and imaged using the water immersion objective. As previously described[14], we experimentally confirmed that the resolution limits of the two approaches were essentially identical.

Image acquisition. During the first imaging session, we selected several regions of brain tissue for longitudinal monitoring across the duration of the time-lapse experiment. Each of these regions contained between 1 and 7 dendritic segments visibly expressing GFP. In each imaging session, we acquired 6–8 image stacks of each selected regions using a voxel size of 0.0725 × 0.0725 × 0.628 μm^3.

Image pre-processing. To improve the visual saliency of fine details within each image stack, initial pre-processing of the image stacks involved a blind deconvolution based on an expectation-maximization routine (Autodeblur from Autoquant). We then aligned all individual images acquired at the same depth in tissue using the TurboReg plug-in routine for ImageJ. Finally, we averaged pixel intensities across the aligned stacks, yielding a single stack that we used in subsequent analyses.

Scoring of dendritic spines. We scored spines using a custom MATLAB interface that supported manual labelling of spines using the computer mouse, measurements of dendrite length and spine position, and alignments of time-lapse sets of image stacks. For each region of tissue monitored, we loaded all the image stacks acquired across time, such that the temporal sequence of the stacks was preserved but the experimenter was blind to their dates of image acquisition during spine scoring. We excluded images whose quality was insufficient to score spines.

We scored spines similarly to as described previously[32] but with a few modifications. We labelled protrusions as dendritic spines only if they extended laterally from the dendritic shaft by >0.4 μm (Extended Data Fig. 6a, b). We did not include protrusions of <0.4 μm in the analysis (Extended Data Fig. 6c, d). When a spine first appeared in the time-lapse image data we assigned it a unique identity. We preserved the spine's identity across consecutive time points if the distances between the spine in question and two or three of its neighbouring spines were stable. In ambiguous cases, which were hardly the norm, we required stability to <2 μm.

The surviving fraction, $S(t)$, at time t was defined as the fraction of spines present on the first imaging day that were also present a time t later. For the deliberate purpose of attaining conservative estimates of (for example, lower bounds on) the proportion of impermanent spines, in the calculation of $S(t)$ we handled the 18% of spines in the recurrent-location category (Fig. 1d and Extended Data Fig. 1d, e) in the following way. When checking pairs of images acquired an interval t apart, we deliberately did not distinguish between whether the second image contained the original spine or its replacement spine at the same location. This approach thereby underestimated spine turnover as inferred from analyses of $S(t)$, implying that our conclusion of CA1 spine impermanence is not only mathematically conservative but also robust to any scoring errors in which we might have erroneously missed a spine that had in fact persisted to subsequent imaging sessions.

Turnover ratio was defined as the sum of spines gained and lost between two consecutive time points normalized by the total number of spines present at these time points. Spines lost or gained were defined as the number of spines lost or gained between two consecutive time points, respectively, normalized by the total number of spines present at these time points.

To make coarse estimates of spine volumes, for stable spines we determined each spine's fluorescence within a manually drawn region of interest (ROI) in the axial section in which the spine head appeared at its biggest diameter. We normalized this value by the fluorescence value attained by moving the ROI to within the nearby dendritic shaft (as in ref. 7).

Statistical analysis. To test for differences in spine densities, turnover ratios and surviving fractions either over time or between different groups, we used non-parametric two-sided statistical testing (Wilcoxon signed-rank, Mann–Whitney *U* and Kruskal–Wallis ANOVA tests) to avoid assumptions of normality and Dunn–Šidák correction for multiple comparisons. Sample size was chosen to match published work[7–9].

To compare experimentally measured spine survival to the theoretical predictions from kinetic modelling, we assessed the goodness-of-fit for each model by using both the reduced chi-squared statistic and the log-likelihood function (Supplementary Methods). Both the mean and covariance of the surviving fraction depended on the model parameters and influenced the goodness-of-fit (Supplementary Methods). We also required that the parameter describing the minimal separation needed to resolve two spines was 0.5–1 μm (other values for this minimal separation are implausible) (Supplementary Methods).

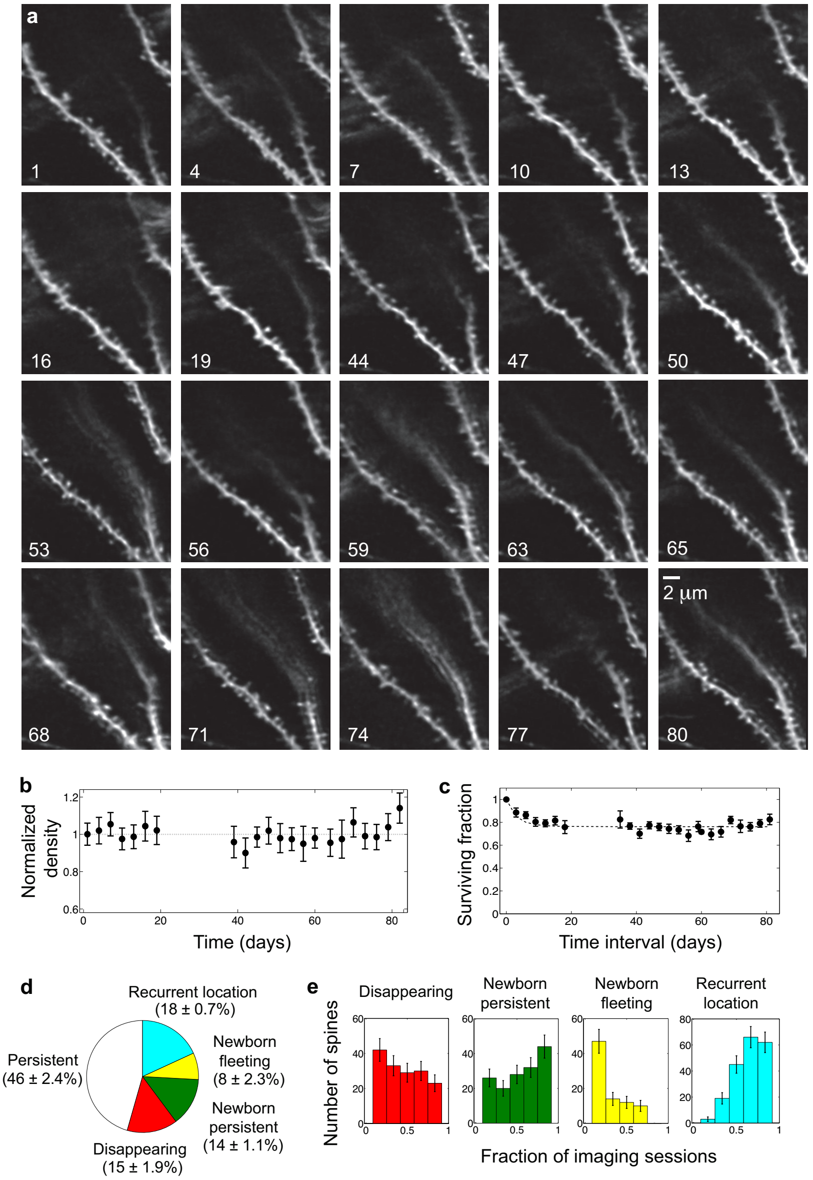
Simulated data sets. We modelled the microscope's optics on the basis of prior measurements[14] and tuned the kinetics of spine turnover, spine geometries and dendrite geometries to produce simulated image sequences that the data analyst judged to be similar to the actual data (Supplementary Information and Extended Data Fig. 5). In some data sets, we matched the simulated spine kinetics to those inferred from our *in vivo* measurements.

31. Feng, G. *et al.* Imaging neuronal subsets in transgenic mice expressing multiple spectral variants of GFP. *Neuron* **28,** 41–51 (2000).

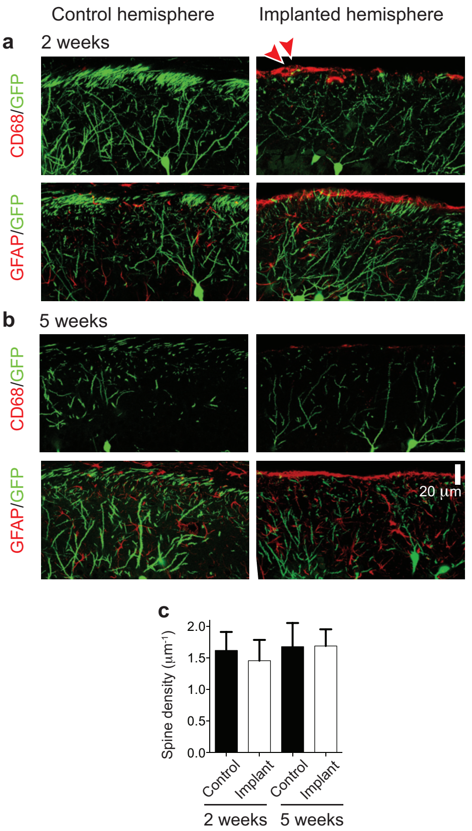
32. Holtmaat, A. *et al.* Long-term, high-resolution imaging in the mouse neocortex through a chronic cranial window. *Nature Protocols* **4,** 1128–1144 (2009).

33. Dombeck, D. A., Harvey, C. D., Tian, L., Looger, L. L. & Tank, D. W. Functional imaging of hippocampal place cells at cellular resolution during virtual navigation. *Nature Neurosci.* **13,** 1433–1440 (2010).

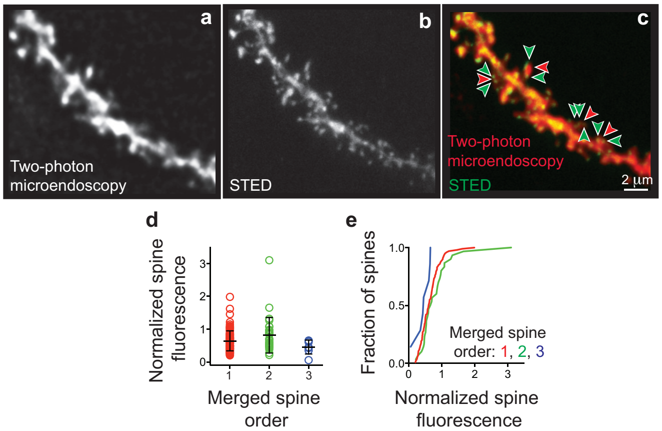
34. Mishchenko, Y. *et al.* Ultrastructural analysis of hippocampal neuropil from the connectomics perspective. *Neuron* **67,** 1009–1020 (2010).



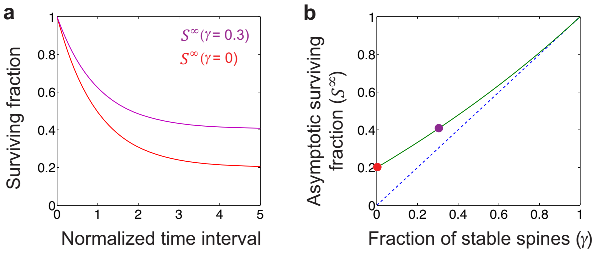
Extended Data Figure 1 | *In vivo* imaging of CA1 spine dynamics over extended time intervals. **a**, *In vivo*, 80-day-long time-lapse image data set sampled at variable intervals. Each image shown is the maximum projection of 4–8 images acquired at adjacent *z*-planes. Scale bar, 2 µm. **b**, **c**, Direct empirical determinations of spine density (**b**) and spine survival (**c**) across the 80 days. A normalized spine density of one corresponds to a measured spine density of 1.16 µm^{-1}. Data points are mean ± s.e.m. for 16 dendrites. **d**, *In vivo*, 22-day-long time-lapse image data set sampled every 3 days. Over 50% of the spines underwent visually noticeable dynamic changes. Pie chart shows the proportions of spines that were persistent or exhibited different patterns of turnover ($n = 1{,}075$ total spines from 4 mice). Colour coding is the same as in Fig. 1d. Error bars are s.e.m. for four mice. **e**, Histograms show the distributions, for each class of spines, of the fraction of imaging sessions in which each spine was observed within the same 22-day data set of **d**. Colour coding is the same as in Fig. 1d. Error bars represent s.d. estimated as counting errors.



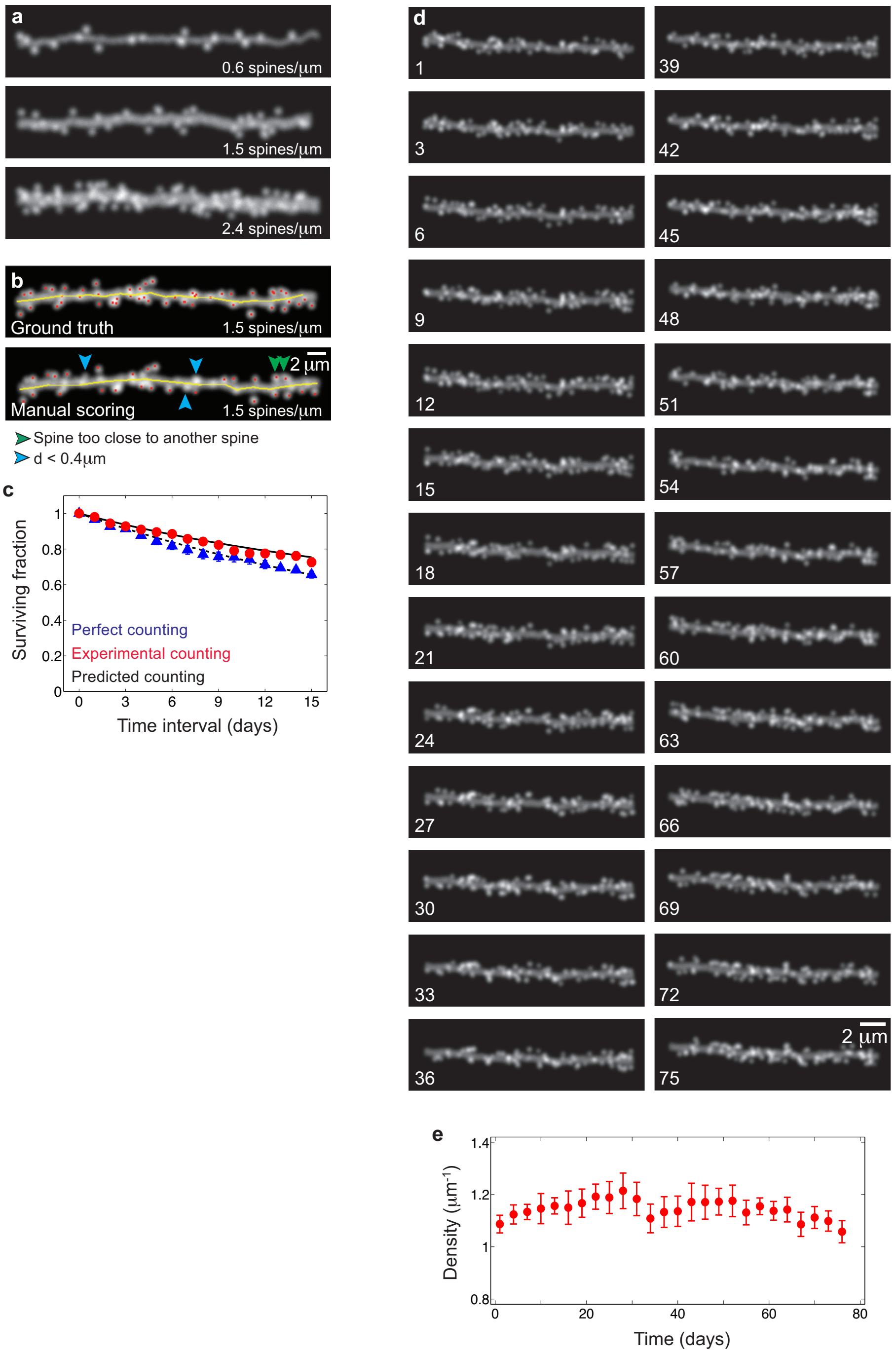
Extended Data Figure 2 | The chronic CA1 preparation induces a minimal, 5-μm-thick layer of glial activation and does not affect spine density. a–c, In two groups of mice, each comprising two *Thy1-GFP* transgenic animals (8–10 weeks old), we implanted the imaging guide tube just dorsal to hippocampal area CA1 following established procedures[4,16,33]. **a**, **b**, We euthanized, sliced and stained the first group after two further weeks (**a**), and the second group after five further weeks (**b**). Confocal fluorescence images of the stained tissue slices revealed activated microglia (CD68 staining, top, red), astrocytes (GFAP staining, bottom, red), GFP-expression pyramidal neurons (green), and permitted quantifications of spine density in CA1 regions both ipsilateral and contralateral to the implant. **a**, Two weeks after implantation, confocal photomicrographs (maximum intensity projection of four separate z-planes, axially spaced 0.2 μm apart) revealed a limited presence of activated microglia (red arrowheads indicate single cells) on the implanted hemisphere (top right) but were virtually non-existent in the contralateral control CA1 from the same mice (top left). Staining for astrocytes on the implanted hemisphere (bottom right) was almost indistinguishable from the control hemisphere (bottom left), except for a 5–10-μm-thick layer of astrocyte label abutting the optical surface of the imaging guide tube. **b**, Five weeks after implantation, confocal photomicrographs (maximum intensity projection of four separate z-planes, axially spaced 0.2 μm apart) revealed an almost undetectable presence of activated microglia on the implanted hemisphere (top right), comparable to the contralateral control hemisphere (top left). Staining for astrocytes in the implanted hemisphere (bottom right) was almost indistinguishable from the control hemisphere (bottom left), except for a 5–10-μm-thick layer of astrocyte label abutting the optical surface of the imaging guide tube. **c**, Mean density of spines on pyramidal cell basal dendrites in the implanted CA1 (white columns) was statistically indistinguishable from the contralateral control CA1 (black columns), at 2 weeks ($P = 0.21$; Mann–Whitney U-test; $n = 11$ dendrites) and at five weeks ($P = 0.98$; Mann–Whitney U-test; $n = 11$ dendrites) after implantation. Error bars are s.d.


a
b
c
Two-photon microendoscopy
STED
Two-photon microendoscopy
STED
2 µm
d
Normalized spine fluorescence
0
1
2
3
Merged spine order
e
Fraction of spines
0.0
0.5
1.0
0
1
2
3
Merged spine order: 1, 2, 3
Normalized spine fluorescence

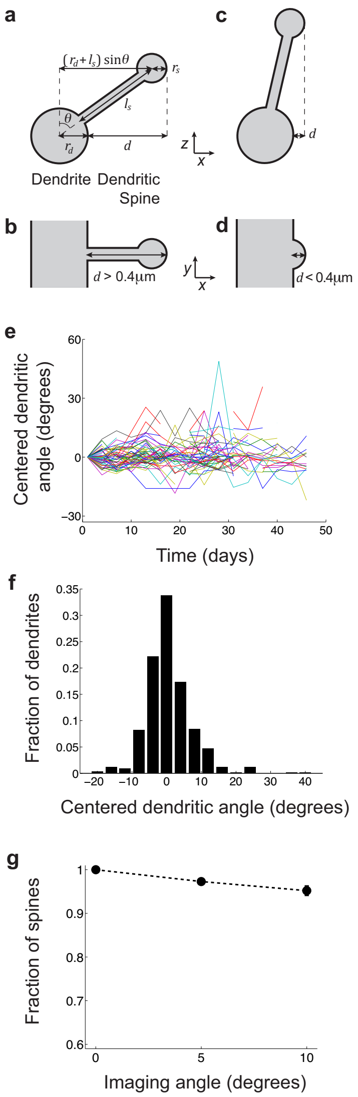
Extended Data Figure 3 | Two-photon and STED imaging of the same CA1 spines in fixed tissue reveals that nearby spines can merge in two-photon images. **a**–**c**, Example two-photon microendoscopy (**a**), STED (**b**), and overlay (**c**) images of the same CA1 basal dendrite, acquired in a fixed brain slice from a *Thy1-GFP* mouse. Nearby spines that are clearly distinguishable in the STED image (green arrowheads) but within the diffraction-resolution limit of two-photon microendoscopy (0.85 NA) appear as single, merged entities within the two-photon image (red arrowheads). The two-photon image shown is the maximum intensity projection of three optical sections axially spaced 0.6 µm apart. The STED image is the maximum intensity projection of six optical sections spaced 0.3 µm apart. Scale, 2 µm. **d**, **e**, To attain an approximate measure of spine volume, we quantified each spine's fluorescence in manually drawn regions of interest (ROIs) and normalized it by the fluorescence value in the nearby dendritic shaft, within an ROI of identical shape and size within a single axial section of the two-photon image stack. To ascertain whether each of the spines scored in the two-photon images was actually a merged spine or not, we consulted the STED images of the same dendrite. Plotted are the normalized fluorescence values (**d**) for unitary spines as well as doublet and triplet merged spines (black lines: mean values ± s.d.; coloured points: data from individual spines), and the cumulative distributions of these measurements (**e**). The distributions of normalized fluorescence were statistically indistinguishable ($P > 0.06$; Kruskal–Wallis ANOVA; $N = 100$, 30 and 7, respectively, for unitary, doublet and triplet spines), probably reflecting the substantial range of CA1 spine geometries (Extended Data Fig. 9).



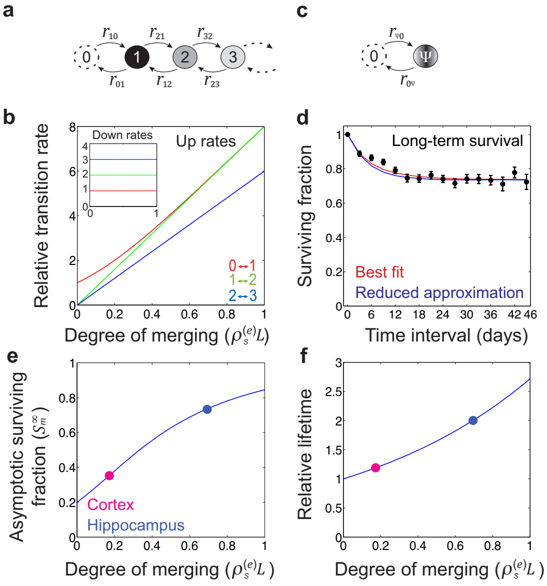
Extended Data Figure 4 | The asymptotic value of the surviving fraction of spines exceeds the fraction of permanent spines. **a**, Surviving fraction curves for models in which the fraction of permanently stable spines is $\gamma = 0$ (red curve) and $\gamma = 0.3$ (purple curve) (Supplementary Methods). The timescale of spine survival was $\tau = 1$, and the filling fraction value was $f = 0.2$. The surviving fraction asymptotes to a value, S^∞, that encodes the fraction of stable spines. (Supplementary Information has a list of all mathematical variables used in this work, and their definitions). **b**, The time asymptotic value of the surviving fraction (green curve) exceeds the fraction of stable spines (dashed blue line). Coloured circles correspond to the surviving fraction curves plotted in **a**.



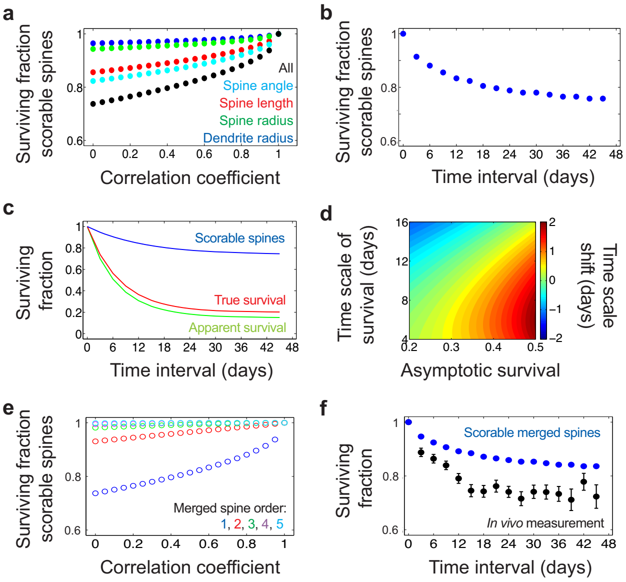
Extended Data Figure 5 | Examples of simulated imaging data sets and their scoring. a, Example simulated images (Supplementary Methods) of dendrites for which the spine density was 0.6 μm^{-1} (top), 1.5 μm^{-1} (middle) and 2.4 μm^{-1} (bottom). **b**, Ground truth and manual scoring of a simulated dendrite for which the spine density was 1.5 μm^{-1}. Green arrowheads indicate counting errors originating from the optical merging of spines (Supplementary Methods). Blue arrowheads indicate counting errors that occur when the spine's projection into the optical plane is too short (Supplementary Methods). Scale bar, 2 μm. **c**, The visually scored surviving fraction (red circles) differed from the true spine surviving fraction (blue triangles), but the departures were well predicted by the kinetic model (solid black curve) (Supplementary Methods). **d**, We simulated and scored a long-term lapse imaging data set with kinetic parameters that matched the best-fit model of Fig. 4b (Supplementary Methods). Even though the data set lacked stable spines, many simulated spines appeared to persist for long time intervals. Scale bar, 2 μm. **e**, Although the spine density in the simulated data was 2.56 μm^{-1}, visual scoring yielded a lower spine density. We used the measured and true spine densities to estimate the extent of merging and the counting resolution (Supplementary Methods). Data points are mean ± s.e.m for 20 simulated dendrites (**c**) or 10 simulated dendrites (**e**).



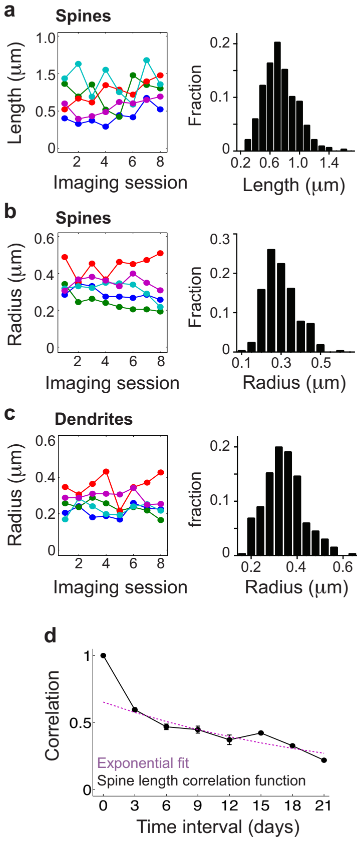
Extended Data Figure 6 | Variability of imaging angles has virtually no impact on determinations of spine turnover. **a**–**g**, We examined empirically whether variations in dendritic angle across different imaging sessions might impact determinations of spine turnover. However, the variations in dendritic angle that were actually present in our data sets were insufficient to cause illusory turnover. **a**, **b**, Dendritic spines can be detected when the angle (θ) between a spine and the normal vector is large (Supplementary Methods). View of a dendrite and spine in the (x, z) (**a**) and optical (x, y) (**b**) planes. **c**, **d**, Dendritic spines cannot be detected when the angle between a spine and the normal vector (θ) is small (Supplementary Methods). View of a dendrite and spine in the (x, z) (**c**) and optical (x, y) (**d**), planes. **e**, For every dendrite and time point, we estimated the dendrite's angle with respect to the optical plane using the three-dimensional coordinates of two manually labelled points on the dendrite chosen to flank the region of dendrite containing the scored spines. Over time, individual dendrites varied about their initial angle ($n = 55$ dendrites tracked over 16 sessions; data set of Fig. 3d). **f**, Distribution of the fluctuations in angle, pooled across the 55 dendrites, relative to the initial angle as seen in the first imaging session. The average magnitude of an angular fluctuation was 4.5°, and 90% of angular fluctuations were $<10°$ in magnitude. Thus, a 5° fluctuation was typical in our data set, whereas a 10° fluctuation was atypically large. **g**, To determine if variability in the imaging angle might impact determinations of spine turnover, we imaged 18 dendrites in fixed slices while deliberately tilting the imaging plane by 0°, 5° and 10°. We made a total of 989 spine observations. Over 95% of spines scored in the 0° condition were also correctly scored when the specimen was tilted by 5° or 10°. Overall, the level of angular fluctuations in the *in vivo* imaging data has virtually no impact on turnover scores.



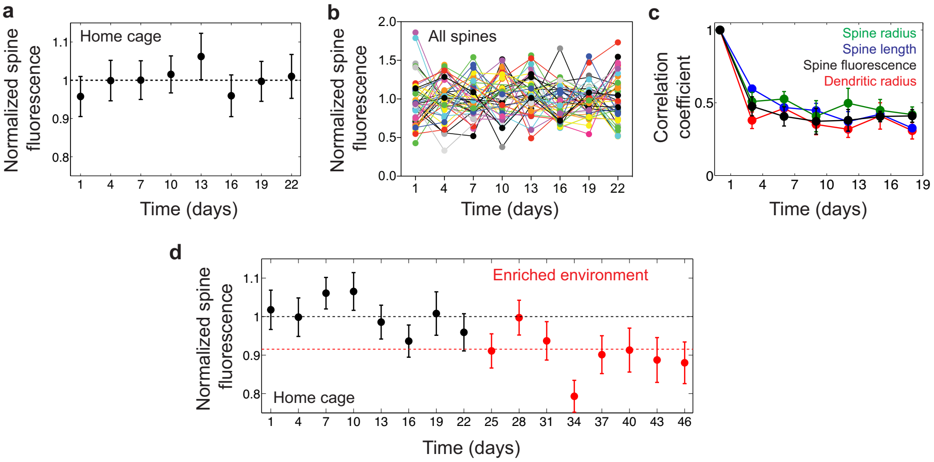
Extended Data Figure 7 | Kinetic modelling well describes how optical merging affects spine turnover dynamics as monitored with finite optical resolution. **a**, Diagram of the kinetic scheme used to describe merged spine dynamics (Supplementary Methods). Each state is labelled by the number of actual spines that have merged in appearance to a single spine (a quantity that we call the merged spine order; Supplementary Methods). State '0' indicates the absence of a spine, and state '1' indicates a spine that is truly unitary. Transitions occur between adjacent states in the kinetic ladder diagram with rate constants r_{mn}. **b**, The rate constants governing increases in merged spine order depend on two parameters (Supplementary Methods): (1) the initial state or merged spine order; and (2) the overall degree of merging in the spine image data set, which is proportional to the product of the spine density and the shortest resolvable interspine interval (denoted L) (Supplementary Methods). By contrast, the rate constants governing decreases in merged spine order (inset) depend only on the initial merged spine order (Supplementary Methods). **c**, In the case when all spines are labile, a collapsed kinetic scheme in which a single state (Ψ) combines all merged spine orders above zero approximates the complete model (**d**) and can be solved mathematically (Supplementary Methods). **d**, The surviving fraction curve generated from the collapsed kinetic scheme (blue curve) fits the empirically observed surviving fraction (black data points) as well as the best-fit model (red). **e**, The asymptotic value of the surviving fraction is a function of the degree of merging (Supplementary Methods). A large degree of merging (as in CA1, blue circle) produces a larger asymptotic value of the merged spine surviving fraction than a small degree of merging (as in the neocortex, purple circle). **f**, The estimated lifetime of merged spines is a function of the degree of merging (Supplementary Methods). A large degree of merging (as in the hippocampal CA1, blue circle) produces a longer relative lifetime of merged spines than a small degree of merging (as in the neocortex, purple circle).



Extended Data Figure 8 | Dynamic spine geometries induce modest levels of apparent spine turnover that cannot explain the turnover measured *in vivo*. **a**–**f**, To study potential effects of fluctuations in spine geometry, we used values for the means and variances of dendrite radius, spine length and spine radius that were determined by electron microscopy[17,34]. We then computationally examined how time-dependent fluctuations in these parameters would affect determinations of spine surviving fraction (Supplementary Methods). **a**, We examined how fluctuations in dendrite radius, spine radius, spine length and spine angle—individually (coloured data points) and all together (black data)—affect the spine surviving fraction when the fluctuating geometric parameters are chosen stochastically in each of two imaging sessions, as a function of the parameter's time correlation between the two sessions (Supplementary Methods). As expected, when the two sessions involved image pairs that were perfectly correlated, the surviving fraction reached 100%. Fluctuations in all four parameters had greater effects than fluctuations in individual geometric parameters. **b**, To estimate the time dependence of the surviving fraction of scorable spines from **a**, we assumed all geometric parameters evolved according to the time-correlation function that we empirically determined from *in vivo* imaging data (Extended Data Fig. 9d). **c**, The apparent surviving fraction is the product of the true surviving fraction and the surviving fraction of scorable spines (Supplementary Methods). For the best-fit kinetic model, the apparent surviving fraction is very close to the true surviving fraction. **d**, The difference between the fitted timescale of the apparent surviving fraction and the true survival timescale is small across the range of model parameters consistent with the *in vivo* data (Supplementary Methods). **e**, The graph plots the lower bound of the surviving fraction of scorable merged spines as a function of the time-correlation function shared by all four geometric parameters, for different merged spine orders. As this lower bound increases rapidly with the merged spine order, artefactual turnover due to unscorable spines is unlikely when spine merging is common (Supplementary Methods). **f**, We combined Fig. 4c and Extended Data Fig. 8e to bound the turnover that could result from unscorable spines (Supplementary Methods). As the empirically measured surviving fraction falls below the lower bound obtained for the surviving fraction of scorable merged spines, ongoing changes in the geometric parameters of spines cannot account for the observed spine turnover.



Extended Data Figure 9 | Dynamics of spine geometries measured *in vivo*. **a**, Time courses of the spine length, measured from the border of the dendritic shaft to the centre of the spine, for five example spines tracked over eight imaging sessions (left). Distribution of spine lengths (right; $n = 344$ spine observations). **b**, Time courses of the spine radius, measured from the border to the centre of the spine, for five example spines tracked over eight imaging sessions (left). Distribution of spine radii (right; $n = 344$ spine observations). **c**, Time courses of the dendritic radius, measured from the border to the centre of the dendrite, at the location of five example spines tracked over eight imaging sessions (left). Distribution of all dendritic radii (right; $n = 344$ dendrite observations). **d**, Experimental spine length time-correlation function and its exponential fit (Supplementary Methods).



Extended Data Figure 10 | Volumes of stable spines fluctuate minimally over time. **a**–**d**, To attain an approximate measure of spine volume for stable spines, we quantified each spine's fluorescence in manually drawn regions of interest (ROIs) and normalized it by the fluorescence value in the nearby dendritic shaft, as determined within a ROI of identical shape and size in a single *z*-section image acquired by two-photon microendoscopy. In addition, each spine's fluorescence value at each time point is normalized to its own mean over the entire experiment. **a**, **b**, Mean (± s.e.m.) fluorescence intensities of all spines (**a**), and individual spines (**b**), from a set of 43 stable spines, across a 21-day span during which mice ($n = 4$) were in their home cages (same data as for Fig. 3a). Dashed black line in **a** indicates the mean over all imaging sessions. **c**, The correlation functions of spine radius (green), length (blue), fluorescence (black) and dendritic radius (red) are indistinguishable from each other. **d**, Mean (± s.e.m.) fluorescence intensities of 61 stable spines across a 46-day span during which mice ($n = 3$) initially resided in their home cages (black data points) but later moved to an enriched environment (red points) (same data set as for Fig. 3d). Dashed black and red lines respectively denote the mean values over the baseline and enriched periods.

doi:10.1038/nature14553

Parent stem cells can serve as niches for their daughter cells

Ana Pardo-Saganta[1,2,3]*, Purushothama Rao Tata[1,2,3]*, Brandon M. Law[1,2,3], Borja Saez[1,3,4], Ryan Dz-Wei Chow[1,2,3], Mythili Prabhu[1,2,3], Thomas Gridley[5] & Jayaraj Rajagopal[1,2,3]

Stem cells integrate inputs from multiple sources. Stem cell niches provide signals that promote stem cell maintenance[1,2], while differentiated daughter cells are known to provide feedback signals to regulate stem cell replication and differentiation[3–6]. Recently, stem cells have been shown to regulate themselves using an autocrine mechanism[7]. The existence of a 'stem cell niche' was first postulated by Schofield in 1978 to define local environments necessary for the maintenance of haematopoietic stem cells[1]. Since then, an increasing body of work has focused on defining stem cell niches[1–6]. Yet little is known about how progenitor cell and differentiated cell numbers and proportions are maintained. In the airway epithelium, basal cells function as stem/progenitor cells that can both self-renew and produce differentiated secretory cells and ciliated cells[8,9]. Secretory cells also act as transit-amplifying cells that eventually differentiate into post-mitotic ciliated cells[9,10] . Here we describe a mode of cell regulation in which adult mammalian stem/progenitor cells relay a forward signal to their own progeny. Surprisingly, this forward signal is shown to be necessary for daughter cell maintenance. Using a combination of cell ablation, lineage tracing and signalling pathway modulation, we show that airway basal stem/progenitor cells continuously supply a Notch ligand to their daughter secretory cells. Without these forward signals, the secretory progenitor cell pool fails to be maintained and secretory cells execute a terminal differentiation program and convert into ciliated cells. Thus, a parent stem/progenitor cell can serve as a functional daughter cell niche.

To establish whether post-mitotic ciliated cells send a conventional feedback signal to regulate the replication of their parent stem and progenitor cells, we genetically ablated ciliated cells using *FOXJ1-creER; LSL-DTA* (Rosa26R-DTA) mice (hereafter referred to as FOXJ1-DTA) (Fig. 1a). Following ciliated cell ablation, the absolute numbers and morphology of secretory progenitor cells (SCGB1A1$^+$) and basal stem/progenitor cells (CK5$^+$) remained unchanged despite the ablation of 78.8% of ciliated cells (on day 5, a total of 24.29 ± 0.3% of all epithelial cells in control mice (identified with the nuclear marker 4′,6-diamidino-2-phenylindole, DAPI$^+$) were FOXJ1$^+$ ciliated cells versus 5.13 ± 0.4% in tamoxifen-treated mice ($n = 3$ mice)) (Fig. 1b, c and Extended Data Fig. 2a, b). Surprisingly, we did not observe the anticipated increase in stem or progenitor cell proliferation and/or their differentiation to replenish missing ciliated cells (Extended Data Fig. 2c–e). Even over extended periods of time, the rates of epithelial proliferation remained similar to those of uninjured controls (Extended Data Fig. 2d). The number of ciliated cells increased at a rate that corresponds to the normal rate of ciliated cell turnover (Fig. 1d). Following ciliated cell ablation, ciliated cell turnover occurs with a half-life of 149 days (Fig. 1e) which mirrors the reported steady-state half-life of approximately 6 months[11]. Additionally, the mesenchymal, haematopoietic, endothelial and smooth muscle cell populations appeared unchanged (Extended Data Fig. 2f, g).

Lacking evidence to support the presence of a feedback mechanism to restore ciliated cell numbers after ablation, we wondered whether basal stem/progenitor cells might regulate secretory daughter cell behaviour by regulating the differentiation of secretory cells into ciliated cells. Thus, we ablated basal cells and simultaneously traced the lineage of secretory progenitor cells using *Scgb1a1-creER;LSL-YFP;CK5-rtTA;tet(O)DTA* mice (hereafter referred to as SCGB1A1-YFP;CK5-DTA), as previously described[12] (Fig. 1f). In addition to the dedifferentiation of secretory cells we previously described following stem cell ablation[12], we observed an increase in lineage-labelled yellow fluorescent protein (YFP$^+$) cells expressing the ciliated cell marker FOXJ1 (8.1 ± 1.6% of YFP$^+$ cells were FOXJ1$^+$ in controls versus 42.4 ± 1.0% in experimental animals) and an accompanying decrease in YFP$^+$ SCGB1A1$^+$ secretory cells (88.5 ± 4% versus 45 ± 3%) ($n = 3$ mice) (Fig. 1g, h). We again observed that ~8% of lineage-labelled secretory cells dedifferentiated into basal cells as previously described[12]. Thus, we can now account for the fates of all lineage-labelled secretory cells after stem cell ablation, as the decrement in secretory cell lineage label (43.5%) is almost precisely equal to the combined increase in lineage-labelled ciliated and basal cells (34% and 8%, respectively). Importantly, lineage-labelled ciliated cells expressed c-MYB, a transcription factor required for ciliogenesis[13,14] and acetylated tubulin (AcTub) confirming that secretory cells differentiated into mature ciliated cells (Extended Data Fig. 3a, b). These results were confirmed by flow cytometry (Extended Data Fig. 3c). In contrast to the changes in the tracheal epithelium in which the total number of ciliated cells increased twofold (625 ± 29 versus 1,208 ± 93 ciliated cells, representing 24.5 ± 1.5% and 61 ± 4.7% of total cells, respectively) (Extended Data Fig. 3d), the underlying mesenchyme remained unchanged in morphology and its complement of haematopoietic, endothelial and smooth muscle cells (Extended Data Fig. 3e, f).

As the Notch pathway has been shown to regulate ciliated versus secretory cell fate choices in the embryonic lung and regenerating adult airway epithelium[15–20], we assessed the expression of Notch pathway components in each cell type of the adult homeostatic airway epithelium. Quantitative real time PCR analysis on purified airway epithelial cells revealed that the Notch1 receptor was highly expressed in basal stem/progenitor cells as previously reported[18], Notch2 and Notch3 were significantly enriched in secretory progenitor cells, and Notch4 was not detected ($n = 3$ mice) (Fig. 2a and Extended Data Fig. 4a).

Signalling through the Notch2 receptor has previously been postulated to regulate secretory cell fate in the embryonic lung[19], in inflammatory cytokine-induced goblet cell metaplasia[20], and we have found it to be activated during secretory cell fate commitment during regeneration[21]. We found that steady-state nuclear Notch2 intracellular domain (N2ICD) expression was restricted to secretory progenitor

[1] Center for Regenerative Medicine, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114, USA. [2] Departments of Internal Medicine and Pediatrics, Pulmonary and Critical Care Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [3] Harvard Stem Cell Institute, Cambridge, Massachusetts 02138, USA. [4] Stem Cell and Regenerative Biology Department, Harvard University, Cambridge, Massachusetts 02138, USA. [5] Center for Molecular Medicine, Maine Medical Center Research Institute, 81 Research Drive, Scarborough, Maine 04074, USA.
*These authors contributed equally to this work.

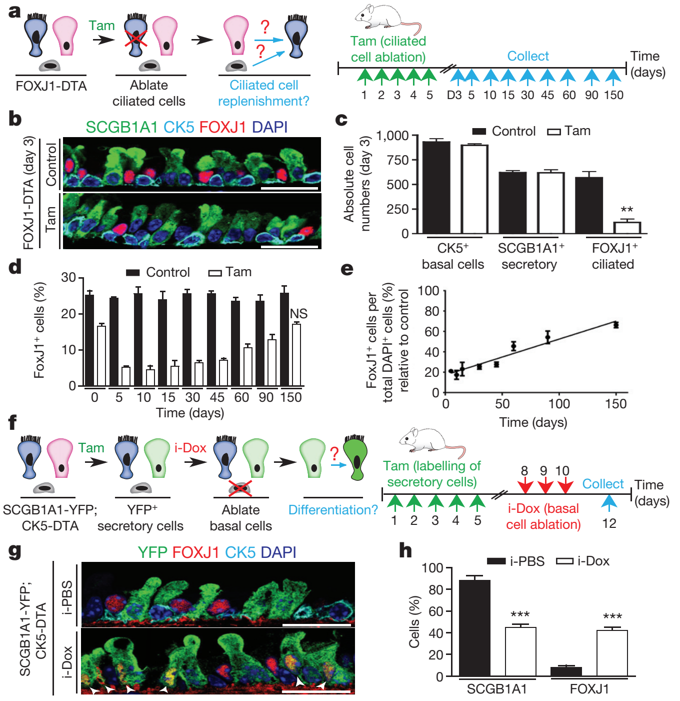


Figure 1 | Secretory progenitor cells differentiate into ciliated cells following basal stem/progenitor cell ablation. **a**, Schematic representation of ciliated cell ablation. Ciliated, secretory and basal cells are shown in blue, pink and grey, respectively. **b**, Immunostaining for SCGB1A1 (green), FOXJ1 (red) and CK5 (cyan) on control (top) or tamoxifen (Tam)-treated FOXJ1-DTA mice (bottom) ($n = 6$ mice). **c**, Absolute cell number of each cell type in both groups ($n = 3$ mice). **d**, Percentage of FOXJ1$^+$ cells per total DAPI$^+$ cells over time ($n = 3$ mice). NS, not significant when compared to day 0 of the same group. **e**, Percentage of FOXJ1$^+$ cells in tamoxifen treated mice ($n = 3$ mice). **f**, Schematic representation of secretory cell lineage labelling and basal cell ablation. **g**, Immunostaining for FOXJ1 (red), YFP (green) and CK5 (cyan) on inhaled (i)-PBS (top) or i-Dox (bottom) treated SCGB1A1-YFP; CK5-DTA mice ($n = 3$ mice). White arrowheads, lineage-labelled ciliated cells. **h**, Percentage of SCGB1A1$^+$ and FOXJ1$^+$ cells per total YFP$^+$ cells. Nuclei, DAPI (4′,6-diamidino-2-phenylindole, blue). n = biological replicates/condition repeated twice (two independent experiments). $^{**}P < 0.01$, $^{***}P < 0.001$. Error bars, means ± s.e.m. Scale bars, 20 μm.

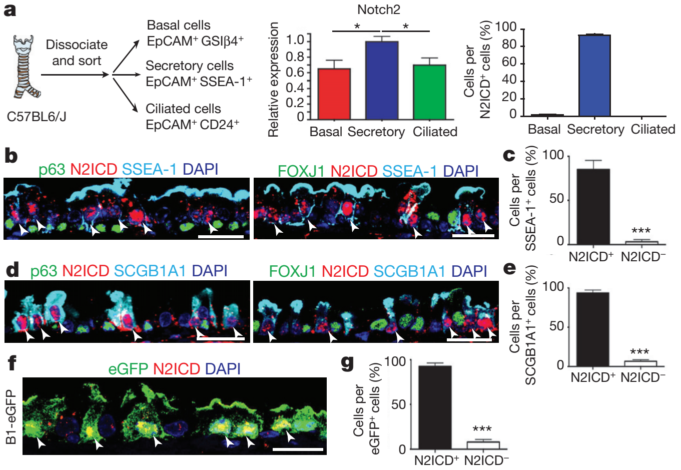


Figure 2 | Secretory progenitor cells show tonic Notch2 activity at steady state. **a**, Schematic representation of airway epithelial cell isolation. Relative messenger RNA expression of *Notch2* in sorted cells ($n = 3$ mice) (middle). Percentage of each cell type per total N2ICD$^+$ cells (right). **b–e**, Immunostaining for p63 (left) or FOXJ1 (right) (green), SSEA-1 (**b**) or SCGB1A1 (**d**) (cyan) and N2ICD (red). Percentage of N2ICD$^+$ cells per total SSEA-1$^+$ (**c**) or SCGB1A1$^+$ (**e**) cells ($n = 3$ mice). **f**, Immunostaining for eGFP (green) and N2ICD (red) in B1-eGFP mice. **g**, Percentage of N2ICD$^+$ cells per total eGFP$^+$ cells ($n = 3$ mice). Nuclei, DAPI (blue). White arrowheads, double-positive cells. Images are representative of $n = 3$ mice (biological replicates). $^*P < 0.05$, $^{***}P < 0.001$. Error bars, means ± s.e.m. Scale bars, 20 μm.

cells (92.7 ± 8% of N2ICD$^+$ cells were secretory cells. $n = 3$ mice), whereas negligible amounts of N2ICD were detected in basal stem/progenitor cells (1.5 ± 3%) and none was seen in ciliated cells (Fig. 2a–e). Consistently, 85.1 ± 5.9% of SSEA-1$^+$ cells and 93.7 ± 2.1% of SCGB1A1$^+$ cells demonstrated active N2ICD expression (Fig. 2b–e). To confirm these observations, we stained airway sections from B1–eGFP mice (in which enhanced GFP (eGFP) is expressed exclusively in secretory cells)[12,22], and found that 92.6 ± 2.2% of eGFP$^+$ cells co-expressed N2ICD (Fig. 2f, g). Unlike the cell specificity associated with N2ICD, we found activated Notch1 (N1ICD) was expressed in most basal stem/progenitor cells and secretory progenitors (Extended Data Fig. 4b). Active N3ICD was detected in subsets of basal, secretory and ciliated cells (Extended Data Fig. 4c). Additionally, the Notch target genes *Hey1* and *HeyL* were enriched in secretory progenitor cells (Extended Data Fig. 4d).

To test directly whether sustained tonic Notch activation is required to maintain secretory cell fate, we abrogated Notch signalling in these cells using *Scgb1a1-creER; LSL-YFP; RBPjk*$^{fl/fl}$ mice (hereafter referred to as SCGB1A1-RBPjk$^{fl/fl}$). The efficient deletion of *RBPjk*, an essential transcription factor required for canonical Notch signalling[23], was confirmed (Extended Data Fig. 5a–c). As a consequence of *RBPjk* deletion, the Notch target genes *Hes1* and *HeyL* were downregulated (Extended Data Fig. 5c). There is a population (approximately 20%) of YFP$^+$ secretory cells in which *RBPjk* deletion has not occurred (yellow arrows in Extended Data Fig. 5a), accounting for the residual *RBPjk* message (Extended Data Fig. 5c). We assessed the fate of lineage-labelled secretory cells following RBPjk loss (Fig. 3a) and found that YFP$^+$ cells were less likely to express secretory cell markers SCGB1A1 (94.4 ± 0.9% versus 31.3 ± 2.2% of YFP$^+$ cells), SCGB3A2 (93.6 ± 1.2% versus 25.7 ± 2.3%) and SSEA-1 (90 ± 1.7% versus 23.5 ± 1%) at the protein level, and were more likely to express the ciliated cell proteins FOXJ1 (5.1 ± 0.6% versus 68.2 ± 3.1%), acetylated tubulin (AcTub) (7.4 ± 1.3% versus 70.6 ± 3.8%) and c-MYB ($n = 6$ mice) (Fig. 3b, c and Extended Data Fig. 5d, e). A decrease in the expression of the secretory cell-specific genes *Scgb1a1* and *Scgb3a2* and an increase in the expression of the ciliated cell genes *FoxJ1* and *c-myb* in lineage-labelled YFP$^+$ cells was also observed ($n = 3$ mice) (Fig. 3d). Similarly, secretory cells that had undergone recombination and lost RBPjk concomitantly lost their characteristic N2ICD expression as they switched fate into FOXJ1$^+$ ciliated cells (Fig. 3e). Less than 0.1% of YFP$^+$ cells co-expressed CK5, suggesting that the lack of Notch signalling in secretory cells is not responsible for the dedifferentiation of secretory cells into basal cells that we previously described following basal cell ablation[12] (Extended Data Fig. 5f, g). The cell fate changes described above were confirmed by flow cytometry (Extended Data Fig. 5h, i) and the phenotype persisted over time (Extended Data Fig. 6a–e). Moreover, overall airway cell proliferation and apoptosis were not affected by RBPjk loss (Extended Data Fig. 6f–k). RBPjk loss induced the direct differentiation of secretory cells into ciliated cells in the absence of proliferation as only 1.7 ± 1.1% of all FOXJ1$^+$ cells had incorporated 5-bromodeoxyuridine (BrdU) over the course of the experiment (Extended Data Fig. 6f) and not a single BrdU$^+$ YFP$^+$ FOXJ1$^+$ ciliated cell was found following continuous BrdU administration (Extended Data Fig. 6h, i). Together these results suggest that tonic canonical Notch activity in secretory progenitor cells is necessary for their continued maintenance at steady-state, and that Notch acts by preventing the differentiation of the secretory progenitor cell pool into the terminally differentiated post-mitotic ciliated cell pool.

To determine whether secretory-cell-specific N2ICD transduces a putative basal cell signal that is required for the maintenance of the

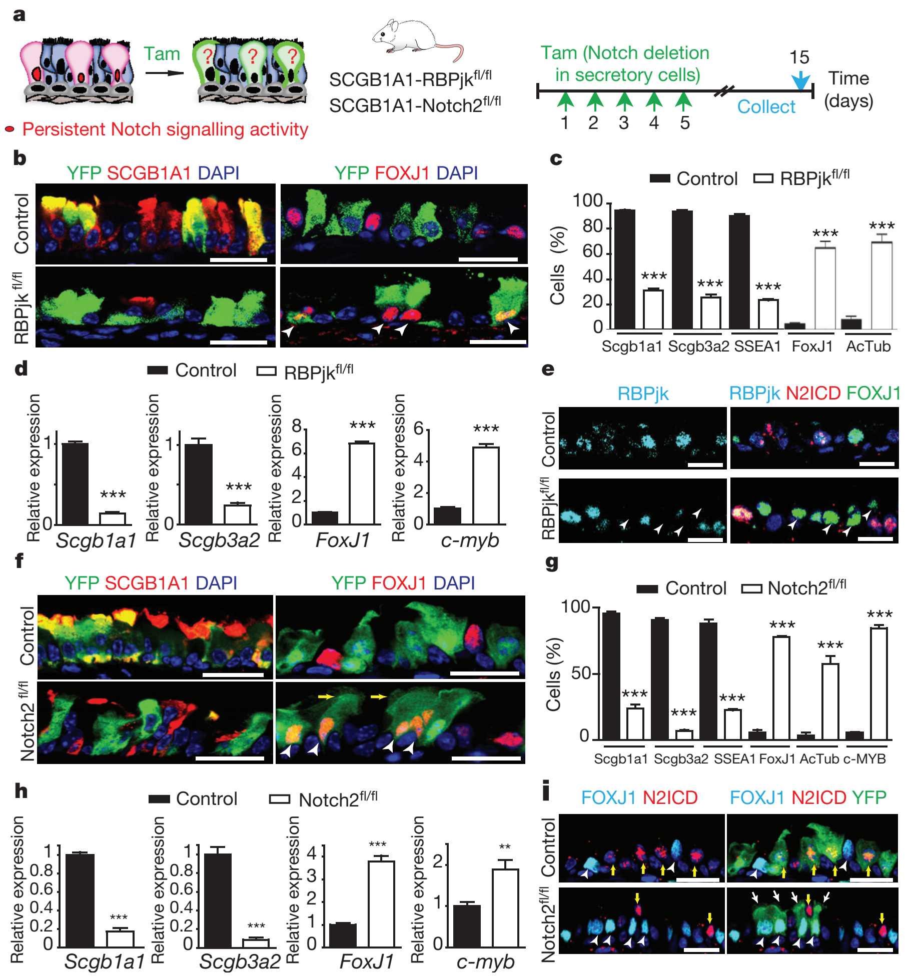


Figure 3 | Tonic Notch2 activity is required to maintain secretory cells by preventing their differentiation into ciliated cells. a, Schematic representation of canonical Notch signalling inhibition in secretory cells. **b**, **f**, Immunostaining for YFP (green) and SCGB1A1 (left) or FOXJ1 (right) (red) in control (top) and experimental (bottom) mice ($n = 6$ mice (**b**); $n = 7$ mice (**f**)). White arrowheads, lineage-labelled ciliated cells. **c**, **g**, Percentage of SCGB1A1$^+$, SCGB3A2$^+$, SSEA-1$^+$, FOXJ1$^+$, AcTub$^+$ and c-MYB$^+$ cells per total YFP$^+$ cells. $n = 3$ mice (**c**); $n = 7$ mice (**g**). **d**, **h**, Relative mRNA expression of *Scgb1a1*, *Scgb3a2*, *FoxJ1* and *c-myb* in control and experimental YFP$^+$ cells ($n = 3$ mice). **e**, Immunostaining for RBPjk (cyan), N2ICD (red) and FOXJ1 (green). White arrowheads, RBPjk$^-$N2ICD$^-$FOXJ1$^+$ cells. **i**, Immunostaining for YFP (green), FOXJ1 (cyan) and N2ICD (red). White arrowheads, FOXJ1$^+$ cells. Yellow arrows, N2ICD$^+$ cells. White arrows, actual cilia in lineage-labelled cells. Nuclei, DAPI (blue). n = biological replicates/condition repeated three times (three independent experiments). ***$P < 0.001$; error bars, means ± s.e.m. Scale bars, 20 μm.

secretory cell pool, we deleted *Notch2* from secretory cells using *Scgb1a1-creER;LSL-YFP;Notch2$^{fl/fl}$* mice (hereafter referred to as SCGB1A1-Notch2$^{fl/fl}$) (Fig. 3a). We first confirmed the efficient deletion of *Notch2* and the downregulation of *Hes1* and *HeyL* (Extended Data Fig. 7a–d). Upon *Notch2* deletion, we observed that lineage-labelled cells ceased to express the secretory cell markers SCGB1A1 (95.6 ± 1.5% versus 23.8 ± 3%), SCGB3A2 (90.8 ± 1.3% versus 6.8 ± 1%) and SSEA-1 (88.2 ± 2.8% versus 22.7 ± 1%) and acquired the expression of the ciliated cell markers FOXJ1 (5.7 ± 2.1% versus 78 ± 0.7%), AcTub (3.7 ± 1.9% versus 57.6 ± 6%) and c-MYB (5.6 ± 0.4% versus 84.5 ± 2.3%) ($n = 7$ mice) (Fig. 3f, g and Extended Data Fig. 7e, f). The expression of secretory cell genes (*Scgb1a1* and *Scgb3a2*) was consistently downregulated in lineage-labelled cells, while ciliated cell genes (*FoxJ1* and *c-myb*) were upregulated ($n = 3$ mice) (Fig. 3h). Intriguingly, YFP staining was present in the actual cilia of lineage-labelled cells, consistent with the terminal differentiation of secretory cells into mature ciliated cells (Fig. 3f, i). Flow cytometry analysis confirmed these cell fate transitions (Extended Data Fig. 7g, h) and also confirmed a lack of dedifferentiation of secretory cells into basal stem cells following Notch pathway modulation (Extended Data Fig. 7i, j). The observation that N2ICD and FOXJ1 expression remained mutually exclusive following *Notch2* deletion also suggested a largely completed cell fate transition (Fig. 3i). However, very rarely, YFP$^+$ cells expressing both markers were observed, leading one to speculate that these rare cells are transient cells caught in the process of differentiating from a secretory cell into a ciliated cell (Extended Data Fig. 8a). Similarly, rare lineage-labelled cells also co-express SSEA-1 and FOXJ1 (Extended Data Fig. 8b). Furthermore, following Notch2 elimination, Ki67 and BrdU incorporation and rates of apoptosis remained unchanged (Extended Data Fig. 8c–g). Additionally, secretory cells directly differentiated into ciliated cells in the absence of proliferation since an insignificant (1.4 ± 1.7%) percentage of FOXJ1$^+$ cells were BrdU$^+$ following continuous BrdU administration (Extended Data Fig. 8d, e). Together these data demonstrate that tonic Notch2 activity within secretory cells is required for the maintenance of secretory cells. Based upon the results of the basal cell ablation, we speculated that the Notch signal-sending cells are basal stem/progenitor cells.

Consistent with prior studies[8,16,18,24], we found that *Dll1* and *Jag2* were expressed in basal stem/progenitor cells, while *Jag1* was enriched in ciliated cells (Fig. 4a), and *Dll3* and *Dll4* were undetectable (data not shown). To remove the putative Notch signal arising from basal stem/progenitor cells, we deleted *Mindbomb1* (Mib1) which is an E3 ubiquitin ligase required for the normal endocytic processing of all Notch ligands[25] in basal cells using *CK5-rtTA; tet(O) Cre; Mindbomb1$^{fl/fl}$* mice (hereafter referred to as CK5-Mib1$^{fl/fl}$) (Fig. 4b). Upon efficient removal of Mib1 (93.3 ± 3.8% of basal cells) (Extended Data Fig. 9a, b), a decrease in SCGB1A1$^+$ (42.8 ± 0.9% versus 26.2 ± 1.0%), SCGB3A2$^+$ (44.6 ± 6.6% versus 6.2 ± 0.7%) and SSEA-1$^+$ secretory cells (49.2 ± 2.6% versus 24.7 ± 1.1%) was accompanied by an increase in FOXJ1$^+$ (30.1 ± 0.9% versus 36.1 ± 1.0%), AcTub$^+$ (21.7 ± 0.7% versus 24.8 ± 0.7%), and c-MYB$^+$ ciliated cells (30.8 ± 2.9% versus 56.2 ± 8.0%) ($n = 4$ mice) (Fig. 4c, d and Extended Data Fig. 9c, d). A corresponding significant decrease in the percentage of N2ICD$^+$ secretory cells was observed (43 ± 1.7% versus 29.6 ± 0.8% of total epithelial cells) (Fig. 4e, f), confirming that Notch ligands emanating from stem cells are necessary for N2ICD activity in secretory cells. These results were confirmed by flow cytometry which additionally revealed that there were no changes in the abundance of basal cells (Extended Data Fig. 9e, f). Rates of proliferation and apoptosis were

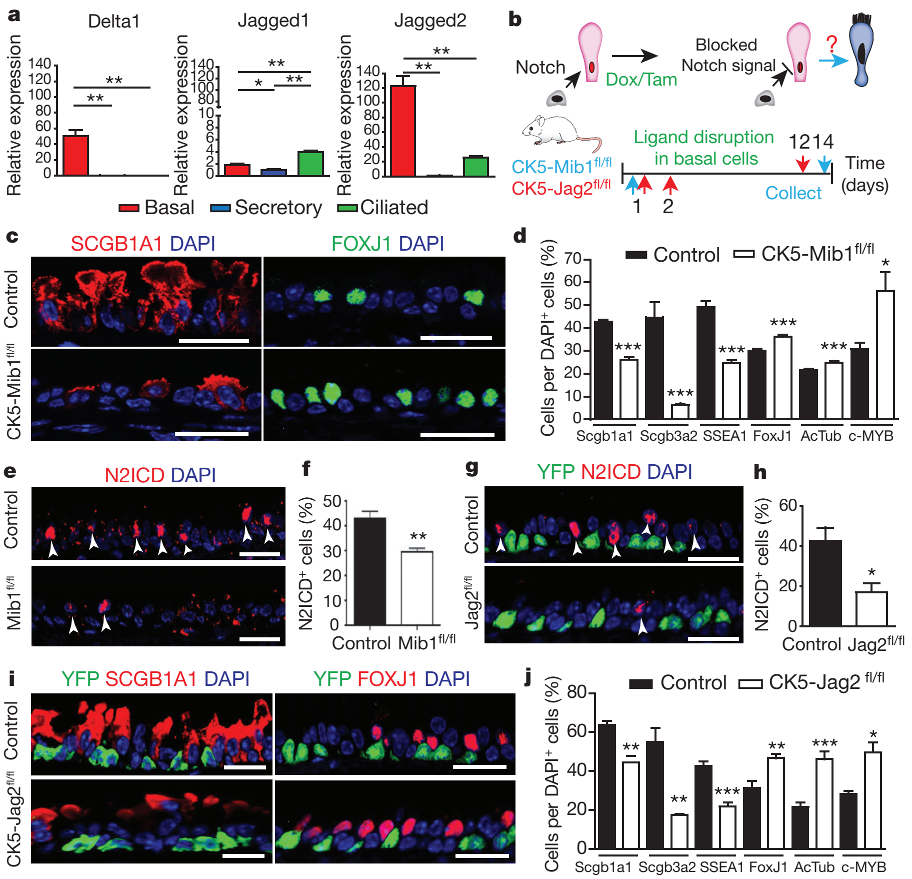


Figure 4 | Basal cell Jagged 2 expression is required to maintain secretory progenitors and prevent their differentiation into ciliated cells. **a**, Relative mRNA expression of Delta1*(Dll1)*, Jagged 1 *(Jag1)* and Jagged 2 *(Jag2)* in sorted cells ($n = 3$ mice). **b**, Schematic representation of Notch ligand disruption in basal cells. **c**, Immunostaining for SCGB1A1 (red, left) and FOXJ1 (green, right) in control (top) and experimental CK5-Mib1$^{fl/fl}$ mice (bottom) ($n = 4$ mice). **d**, **j**, Percentage of SCGB1A1$^+$, SCGB3A2$^+$, SSEA-1$^+$, FOXJ1$^+$, AcTub$^+$ and c-MYB$^+$ cells in control and experimental mice. $n = 4$ mice (**d**); $n = 5$ mice (**j**). **e**, **g**, Immunostaining for N2ICD (red) and YFP (green, in **g**) $n = 4$ mice (**e**); $n = 5$ mice (**g**). White arrowheads, N2ICD$^+$ cells. **f**, **h**, Percentage of N2ICD$^+$ cells per total DAPI$^+$ cells ($n = 4$ mice (**f**); $n = 5$ mice (**h**)). **i**, Immunostaining for YFP (green) and SCGB1A1 (left) or FOXJ1 (right) (red) in control (top) and experimental CK5-Jag2$^{fl/fl}$ mice (bottom) ($n = 5$). Nuclei, DAPI (blue). n = biological replicates/condition repeated twice (CK5-Mib1 mice) or three times (CK5-Jag2 mice). $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. Error bars, means ± s.e.m. Scale bar, 20 μm.

also unchanged (Extended Data Fig. 9g–l) and a negligible amount (0.77 ± 1.5%) of FOXJ1$^+$ cells were found to incorporate BrdU after continuous BrdU administration (Extended Data Fig. 9i, j). In addition, the cell fate changes described above continued to be present 5 weeks after *Mib1* deletion (Extended Data Fig. 9m).

These results are consistent with the model that basal stem/progenitor cells send an essential signal to secretory progenitor cells, and this signal is necessary for the maintenance of the appropriate balance of cell types in the airway epithelium. As Jag2 is the most abundantly expressed ligand in basal stem cells (Fig. 4a), we knocked down *Jag2* expression *in vitro* using short hairpin RNA (shRNA) lentiviral vectors (Extended Data Fig. 10a–c). This resulted in a decrease in *Scgb1a1* and *Scgb3a2* expression and an increase in *FoxJ1* and *c-myb* expression (Extended Data Fig. 10d), resembling the effects of *in vivo* Notch signalling disruption. To confirm that Jag2 is the signal emanating from basal stem/progenitor cells, we generated *CK5-creER; LSL-YFP; Jagged2*$^{fl/fl}$ mice (hereafter referred to as CK5-Jag2$^{fl/fl}$) to genetically remove Jag2 from basal stem/progenitor cells *in vivo* (Fig. 4a). *Jag2* deletion was confirmed (Extended Data Fig. 10e) and although the efficiency of recombination as judged by the number of YFP$^+$ recombined cells was approximately 10% (Extended Data Fig. 10f), the deletion caused a striking decrease in N2ICD$^+$ suprabasal cells (43 ± 6.6% versus 17 ± 4.5% of total airway epithelial cells) (Fig. 4g, h) confirming that Jag2 is the basal cell signal responsible for activating N2ICD in secretory cells. We observed a consistent decrease in SCGB1A1$^+$ (63 ± 2.1% versus 44.4 ± 3.3%), SCGB3A2$^+$ (55 ± 7% versus 17.5 ± 0.5%) and SSEA-1$^+$ secretory cells (42.8 ± 2% versus 21.8 ± 2%) and a concomitant increase in FOXJ1$^+$ (31.3 ± 3.6% versus 46.6 ± 2.2%), AcTub$^+$ (21.7 ± 2.1% versus 46.2 ± 3.9%) and c-MYB$^+$ ciliated cells (28.2 ± 2.1% versus 49.6 ± 11.3%) ($n = 5$ mice) (Fig. 4i, j and Extended Data Fig. 10g, h). Results were also confirmed by flow cytometry (Extended Data Fig. 10i, j). Furthermore, we found no difference in the percentage of p63$^+$ (also known as Trp63) basal cells (Extended Data Fig. 10k, l). N2ICD and FOXJ1 expression was mutually exclusive, consistent with a completed cell fate transition (Extended Data Fig. 10m), and there were no differences in overall proliferation and apoptosis (Extended Data Fig. 10n–r).

Together, our results show that basal stem/progenitor cells regulate the maintenance of their own progeny through a mechanism in which basal-stem-cell-produced Jag2 activates Notch2 in daughter secretory progenitor cells to prevent secretory cell differentiation into post-mitotic ciliated cells (Extended Data Fig. 1).

Schofield first introduced the term niche to make sense of experimental evidence that suggested the presence of local environments necessary for the maintenance of haematopoietic stem cells[1]. However, he was explicit in referring to stem cell niches. We now show that stem/progenitor cells themselves serve as 'daughter cell niches' (Extended Data Fig. 1c). We would like to suggest that reciprocal forms of niche-type regulation may be a general feature of many tissues in which stem, progenitor and differentiated cells might all regulate the maintenance of one another.

To serve as a progenitor cell niche, airway stem/progenitor cells use a 'forward signal' sent to their own progeny. We define a forward signal as a signal that is relayed from a parent cell to its daughter cell. Interestingly, in parallel to our mammalian example, in the fly midgut, a forward Notch signal is sent from an intestinal stem cell to alter the fate choice of its own downstream progeny[26]. However, from one setting to the next, Notch, with its myriad receptors and ligands, will inevitably be deployed in very divergent ways, even within the same tissue[23,24,27,28]. For example, following injury, airway basal stem/progenitor cells use a mechanism akin to lateral inhibition to segregate their lineages[21], whereas pan-epithelial *Jag2* deletion alters the distribution of airway progenitors in the embryonic airway epithelium, and in this context Notch3 is thought to be the relevant receptor[24]. Notably, we identify Notch2 as the receiving receptor on secretory cells. N2ICD is, to the best of our knowledge, the first transcription factor that has been found to be specific to steady-state adult airway secretory progenitor cells.

More generally, we note that differentiated cells are commonly thought to send back signals to their respective stem and progenitor cells to regulate their proliferation and differentiation[3–6]. This process is generally termed feedback regulation, and we were surprised not to see evidence of such a regulatory mechanism following ciliated cell ablation. More recently, self signals have been identified that mediate autocrine stem cell regulation[7]. As we demonstrate the existence of a forward signal, we would like to suggest that 'forward regulation' by stem cells is likely to exist (Extended Data Fig. 1d). Although it is tempting to call this form of regulation 'feed-forward regulation' to contrast it with 'feedback regulation', this term has been used in control theory to denote a more complex form of regulation that involves three discrete entities that interact in a loop[29,30]. Therefore, we opt to propose the simpler term 'forward regulation'. To illustrate what we intend to suggest, we note that Notch signals in fly intestinal stem cells occur at varying levels of Notch activation that in turn determine daughter cell fate[26]. Thus, the regulation of these forward Notch signals could be used to alter the distribution and ratio of daughter cell types. In our case, perhaps fluctuations in basal cell ligand levels determine the rate of ciliated cell turnover? And how would such forward signals be modulated following tissue injury? A recent study points to Notch2 as a receptor relevant to human asthma[20]. Perhaps increasing basal cell ligand concentration is a mechanism used to engender the asthmatic epithelial phenotype in which secretory daughter cells differentiate into mucous-secreting goblet cells. Thus, we speculate that stem cells, using forward regulatory mechanisms, may orchestrate many tissue-wide changes, rather than merely acting as a source of new cells.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 December 2014; accepted 1 May 2015.

Published online 6 July 2015.

1. Schofield, R. The relationship between the spleen colony-forming cell and the haemopoietic stem cell. *Blood Cells* **4,** 7–25 (1978).
2. Scadden, D. T. The stem-cell niche as an entity of action. *Nature* **441,** 1075–1079 (2006).
3. Hsu, Y. C. & Fuchs, E. A family business: stem cell progeny join the niche to regulate homeostasis. *Nature Rev. Mol. Cell Biol.* **13,** 103–114 (2012).
4. Bruns, I. *et al.* Megakaryocytes regulate hematopoietic stem cell quiescence through CXCL4 secretion. *Nature Med.* **20,** 1315–1320 (2014).
5. Hsu, Y. C., Li, L. & Fuchs, E. Transit-amplifying cells orchestrate stem cell activity and tissue regeneration. *Cell* **157,** 935–949 (2014).
6. Sato, T. *et al.* Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. *Nature* **469,** 415–418 (2011).
7. Lim, X. *et al.* Interfollicular epidermal stem cells self-renew via autocrine Wnt signaling. *Science* **342,** 1226–1230 (2013).
8. Rock, J. R. *et al.* Basal cells as stem cells of the mouse trachea and human airway epithelium. *Proc. Natl Acad. Sci. USA* **106,** 12771–12775 (2009).
9. Rock, J. R. & Hogan, B. L. M. Epithelial progenitor cells in lung development, maintenance, repair, and disease. *Annu. Rev. Cell Dev. Biol.* **27,** 493–512 (2011).
10. Rawlins, E. L. *et al.* The role of Scgb1a1$^+$ Clara cells in the long-term maintenance and repair of lung airway, but not alveolar, epithelium. *Cell Stem Cell* **4,** 525–534 (2009).
11. Rawlins, E. L. & Hogan, B. L. M. Ciliated epithelial cell lifespan in the mouse trachea and lung. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **295,** L231–L234 (2008).
12. Tata, P. R. *et al.* Dedifferentiation of committed epithelial cells into stem cells *in vivo. Nature* **503,** 218–223 (2013).
13. Pan, J. H. *et al.* Myb permits multilineage airway epithelial cell differentiation. *Stem Cells* **32,** 3245–3256 (2014).
14. Tan, F. E. *et al.* Myb promotes centriole amplification and later steps of the multiciliogenesis program. *Development* **140,** 4277–4286 (2013).
15. Tsao, P. N. *et al.* Notch signaling controls the balance of ciliated and secretory cell fates in developing airways. *Development* **136,** 2297–2307 (2009).
16. Morimoto, M. *et al.* Canonical Notch signaling in the developing lung is required for determination of arterial smooth muscle cells and selection of Clara versus ciliated cell fate. *J. Cell Sci.* **123,** 213–224 (2010).
17. Guseh, J. S. *et al.* Notch signaling promotes airway mucous metaplasia and inhibits alveolar development. *Development* **136,** 1751–1759 (2009).
18. Rock, J. R. *et al.* Notch-dependent differentiation of adult airway basal stem cells. *Cell Stem Cell* **8,** 639–648 (2011).
19. Morimoto, M., Nishinakamura, R., Saga, Y. & Kopan, R. Different assemblies of Notch receptors coordinate the distribution of the major bronchial Clara, ciliated and neuroendocrine cells. *Development* **139,** 4365–4373 (2012).
20. Danahay, H. *et al.* Notch2 is required for inflammatory cytokine-driven goblet cell metaplasia in the lung. *Cell Rep.* **10,** 239–252 (2015).
21. Pardo-Saganta, A. *et al.* Injury induces direct lineage segregation of functionally distinct airway basal stem/progenitor cell subpopulations. *Cell Stem Cell* **16,** 184–197 (2015).
22. Miller, R. L. *et al.* V-ATPase B1-subunit promoter drives expression of EGFP in intercalated cells of kidney, clear cells of epididymis and airway cells of lung in transgenic mice. *Am. J. Physiol. Cell Physiol.* **228,** C1134–C1144 (2005).
23. Kopan, R. & Ilagan, M. X. G. The canonical Notch signaling pathway: unfolding the activation mechanism. *Cell* **137,** 216–233 (2009).
24. Mori, M. *et al.* Notch3-Jagged signaling controls the pool of undifferentiated airway progenitors. *Development* **142,** 258–267 (2015).
25. Koo, B. K. *et al.* An obligatory role of Mind bomb-1 in Notch signaling of mammalian development. *PLoS ONE* **2,** e1221 (2007).
26. Ohlstein, B. & Spradling, A. Multipotent *Drosophila* intestinal stem cells specify daughter cell fates by differential Notch signaling. *Science* **315,** 988–992 (2007).
27. Stamataki, D. *et al.* Delta1 expression, cell cycle exit, and commitment to a specific secretory fate coincide within a few hours in the mouse intestinal stem cell system. *PLoS ONE* **6,** e24484 (2011).
28. Ambler, C. A. & Watt, F. M. Adult epidermal Notch activity induces dermal accumulation of T cells and neural crest derivatives through upregulation of jagged 1. *Development* **137,** 3569–3579 (2010).
29. Mangan, S. & Alon, U. Structure and function of the feed-forward loop network motif. *Proc. Natl Acad. Sci. USA* **100,** 11980–11985 (2003).
30. Alon, U. *An Introduction to Systems Biology: Design Principles of Biological Circuits* (Chapman & Hall/CRC, 2007).

Acknowledgements We thank A. Glick for providing the *CK5rtTA* mice, B. Hogan for providing *CK5creER* and *SCGB1A1creER* mice and Y.-Y. Kong for sharing the *Mib1* floxed mice. B. Z. Stanger provided the *RBPjk* floxed mice and shared protocols for the immunohistochemical detection of Notch components. We also thank B. Stripp for providing the goat anti-SCGB1A1 antibody. We wish to extend our thanks to all of the members of the Rajagopal laboratory and the HSCI flow cytometry core facility. This research was supported by the New York Stem Cell Foundation (J.R. is a New York Stem Cell Foundation-Robertson Investigator), by a National Institutes of Health-National Heart, Lung, and Blood Institute Early Career Research New Faculty (P30) award (5P30HL101287-02), an RO1 (RO1HL118185) from NIH-NHLBI (to J.R.) and a Harvard Stem Cell Institute (HSCI) Junior Investigator Grant (to J.R.). J.R. is also the Maroni Research Scholar at Massachusetts General Hospital.

Author Contributions A.P.-S. designed and performed the experiments and co-wrote the manuscript; P.R.T. performed the ablation experiments and edited the manuscript; B.M.L. optimized the immunodetection of N2ICD, analysed the phenotype of *RBPjk* and *Mib1* deletion *in vivo* and co-wrote the manuscript; B.S. performed flow cytometry experiments and analysis, contributed to the *in vitro* experiments and edited the manuscript; R.D.-W.C. and M.P. helped with the analysis of the *in vivo* experiments; T.G. contributed to Jag2 deletion *in vivo* experiments; J.R. suggested and co-designed the study and co-wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.R. (jrajagopal@mgh.harvard.edu).

METHODS

Animals. *FOXJ1-creER*[11], *CK5-rtTA*[31], *Scgb1a1-creER*[10], *tet(O)cre* (JAX 006224), *CK5-creER*[32], *tet(O)DTA*[33], *Rosa26R-DTA* (JAX 009669), *RBPjk*$^{fl/fl}$[34], *LSL-YFP* (JAX 006148), *Mindbomb1*$^{fl/fl}$[25], *Notch2*$^{fl/fl}$ (JAX 010525), *Jag2*$^{fl/fl}$[35] and C57BL6/J (JAX 000664) mice were previously described. Progeny of *Scgb1a1-creER* and *LSL-YFP* crosses as well as *CK5-rtTA* and *tet(O)DTA* crosses were subsequently mated to generate *Scgb1a1-creER;LSL-YFP;CK5-rtTA;tet(O)DTA* mice[12]. These mice were treated with tamoxifen and then with inhaled PBS (control) or inhaled doxycycline as previously described[12]. *Scgb1a1-creER* mice were crossed with *RBPjk*$^{fl/fl}$ mice to generate secretory-progenitor-specific *Scgb1a1-creER;RBPjk*$^{fl/fl}$ conditional knockout mice. To allow for lineage tracing, these mice were crossed with *LSL-YFP* mice to generate *Scgb1a1-creER;LSL-YFP;RBPjk*$^{fl/fl}$ mice. Tamoxifen was administered by intraperitoneal injection (2 mg per day) for five consecutive days to induce the Cre-mediated recombination. Similarly, *Scgb1a1-creER;LSL-YFP;Notch2*$^{fl/fl}$ mice were generated and treated. *CK5-rtTA* and *tet(O)cre* mice were crossed to generate *CK5-rtTA;tet(O)cre* mice. *CK5-rtTA;tet(O)cre* mice were crossed with *Mindbomb1*$^{fl/fl}$ mice to generate basal-stem-cell-specific *CK5-rtTA;tet(O)Cre;Mindbomb1*$^{fl/fl}$ conditional knockout mice. Doxycycline administration was performed through drinking water (1 mg per ml) for 2 weeks as described previously[21,36]. *CK5-creER;LSL-YFP;Jag2*$^{fl/fl}$ mice were generated and treated, in this case with 2 doses of tamoxifen, due to a higher sensitivity of this strain to the compound. Mice were euthanized 10 days after the last tamoxifen injection. Male 6–12-week–old mice were used for experiments except in specific circumstances in which breeding limitations led to the use of females in the following strains: *Scgb1a1-creER;LSL-YFP;CK5-rtTA;tet(O)DTA* and *CK5-rtTA;tet(O)Cre;Mindbomb1*$^{fl/fl}$ mice. Similarly aged mice were used for both control and treated animals. Controls include corn oil-treated *FOXJ1-creER*; *tet(O)DTA* mice, i-PBS treated Tam-induced *Scgb1a1-creER;LSL-YFP;CK5-rtTA;tet(O)DTA* mice, Tam-treated *Scgb1a1-creER;LSL-YFP;RBPjk*$^{fl/+}$ mice, Tam-treated *Scgb1a1-creER;LSL-YFP;RBPjk*$^{+/+}$ mice, Tam-treated *Scgb1a1-creER;LSL-YFP;Notch2*$^{+/+}$ mice, doxycycline-treated *CK5-rtTA;tet(O)Cre;Mib1*$^{+/+}$ mice and Tam-treated *CK5-creER;LSL-YFP;Jag2*$^{+/+}$ mice. BrdU (5 mg) 2 h before euthanasia in all cases. Additionally, we treated mice with 1 mg ml^{-1} of BrdU in drinking water from the time of the last tamoxifen injection to euthanisia to analyse proliferative events occurring as a consequence of genetic modulation. We analysed at least 3–7 mice per condition in each experiment and all the experiments were repeated at least three times with the exception of *CK5-rtTA;tet(O)Cre;Mindbomb1*$^{fl/fl}$ and the cell ablation experiments that were repeated twice. All procedures and protocols were approved by the MGH Subcommittee on Research Animal Care in accordance with NIH guidelines.

Tissue preparation, immunohistochemistry and immunofluorescence. Mouse trachea were removed using sterile technique and then fixed in 4% paraformaldehyde for 2 h at 4 °C, washed with PBS, and transferred to a 30% sucrose solution overnight. For immunofluorescence, airways were embedded in OCT and cryosectioned as transverse 7-μm sections. Cryosections were stained with the previously described protocol[12,21,36,37]. The following antibodies were used: rabbit anti-caspase3, cleaved (1:100, 9661, Cell Signaling); rabbit anti-cytokeratin 5 (1:1000; ab53121, Abcam); mouse anti-FOXJ1 (1:500; 14-9965, eBioscience); chicken anti-green fluorescent protein (1:500; GFP-1020, Aves Labs); goat anti-GFP (1:100; NB-100-1770, Novus Biologicals); anti-Ki67 (1:200; ab15580, Abcam); rat anti-RBPjk (1:100; SIM-2ZRBP2, Cosmobio); goat anti-SCGB1A1 (1:500; provided by B. Stripp); goat anti-CC10 (1:100; sc-9772, Santa Cruz Biotechnology), rabbit anti-SCGB3A2 (1:100; provided by Shioko Kimura); mouse anti-p63 (1:100; sc-56188, Santa Cruz Biotechnology); mouse IgM anti-SSEA-1 (1:100; 14-8813-82, eBioscience), mouse anti-tubulin, acetylated (1:100; T6793, Sigma), rabbit anti-alpha smooth muscle actin (1:100; ab5694, Abcam), rat anti-CD45 (1:100; 14-0451, eBioscience) and rat anti-CD31 (1:100; 553370, BD Pharmingen). BrdU incorporation was detected using Amersham Cell Proliferation Kit (RPN20, GE Healthcare, Waukesha, WI). Cell death was detected using DeadEnd Fluorometric TUNEL System (G3250, Promega, Madison, WI). Appropriate secondary antibodies (Life Technologies' Alexa Fluor series 488, 594 or 647) were diluted 1:500. In the case of rabbit anti-Notch2 (1:2000; D67C8, Cell Signaling), rabbit anti-activated Notch1 (1:1500, ab8925, Abcam), rabbit anti-Notch3 (1:1500, sc-5593, Santa Cruz Biotechnologies), rabbit anti-c-MYB (1:3000; sc-519, Santa Cruz Biotechnology) and rabbit anti-Mindbomb1 (1:500, M6073, Sigma), following primary antibody incubation, sections were washed and incubated with anti-Rabbit-HRP conjugate (1:1,000; 170-6514, Bio-Rad) for 1 h at room temperature followed by tyramide signal amplification. Sections were then washed an incubated for 30 min at room temperature with streptavidin-594 (1:1,000; S-11227, Life Technologies)[21]. For more information on the protocol to detect low levels of c-MYB and N2ICD using tyramide signalling amplification, please refer to the Rajagopal laboratory website (http://www.massgeneral.org/regenmed/staff/Rajagopallab).

Microscopy and imaging. Tissue was imaged using an Olympus FluoView FV10i confocal microscope (Olympus Corporation). Cells were manually counted based on immunofluorescence staining of markers for each of the respective cell types[21,37]. Briefly, cell counting was performed on the basis of nuclear staining with DAPI (nuclei) and specific cell markers. Cells were counted using ×40 magnification fields (each field represented 250 μm of epithelium) covering the whole tracheal epithelium, from cartilage ring 1 to 10, of each mouse. This includes approximately 1,300 to 1,800 DAPI^{+} cells per experiment. In *CK5-creER;LSL-YFP;Jag2*$^{fl/fl}$ mice, given the low (approximately 10%) rate of genetic recombination, we showed images in regions where there were patches of YFP^{+} basal cells that had undergone recombination, and therefore *Jag2* deletion. Of note, cell counts were performed manually throughout the entire tracheal epithelium, and were not restricted to areas of basal cell recombination even in these mice. Images were processed and analysed using ImageJ/Fiji (NIH) and Adobe Photoshop Creative Suite 5 (Adobe).

Cell dissociation, FACS and flow cytometry analysis. Airway epithelial cells from trachea were dissociated using papain solution as previously described[37]. Briefly, following trachea removal, airway tissue was cut into small fragments and transferred to a 2 ml solution containing 1 ml 100 U of pre-activated papain (Worthington Biochemical Corporation, catalogue number LK003182) and 1 ml of activation buffer as per the manufacturer's protocol. Tissue fragments were incubated on a shaking platform for 90 min at 37 °C. The cell suspension was passed through a 70 μm cell strainer to remove airway husks and pelleted for 5 min at 400*g*. The supernatant was aspirated and the pellet was resuspended in ovomucoid solution (Worthington Biochemical Corporation, catalogue number LK003182) for 20 min at 4 °C to inactivate residual papain activity. Dissociated cells were stained with the following antibodies: EpCAM-PECy7 (1:50; 25-5791-80, eBiosciences) or EpCAM-APC (1:50; 17-5791, eBiosciences); GSIβ4 (Griffonia Simplicifolia Isolectin beta4)-Biotin (L2120, Sigma); SSEA-1 eFluor 650NC (1:75, 95-8813-41, eBiosciences); CD24-PE (1:100, 553262, BD Pharmingen). Primary antibodies were incubated for 30 min in 2.5% FBS in PBS on ice. FACS and flow cytometry was performed on a BD FACSAria II sorter at the CRM Flow Cytometry Core (Boston, MA). All aforementioned cell sortings were previously gated for EpCAM to exclusively select epithelial cells. Of note, differences in the percentage of each airway epithelial cell type analysed by flow cytometry might differ from the quantitation performed by cell counting. This reflects the use of cell surface markers for flow analysis (CD24 for ciliated cells) in contrast to cell counts based on the nuclear transcription factors (such as FoxJ1 and c-MYB for ciliated cells). Additionally, flow cytometry involves enzymatic tracheal dissociation and cells may die in this process and some cell types might demonstrate differential viability following enzymatic dissociation. Sorted cells were lysed immediately in TRI Reagent (Sigma) and RNA was extracted as previously described[37]. Data were analysed on FlowJo Software (version 10).

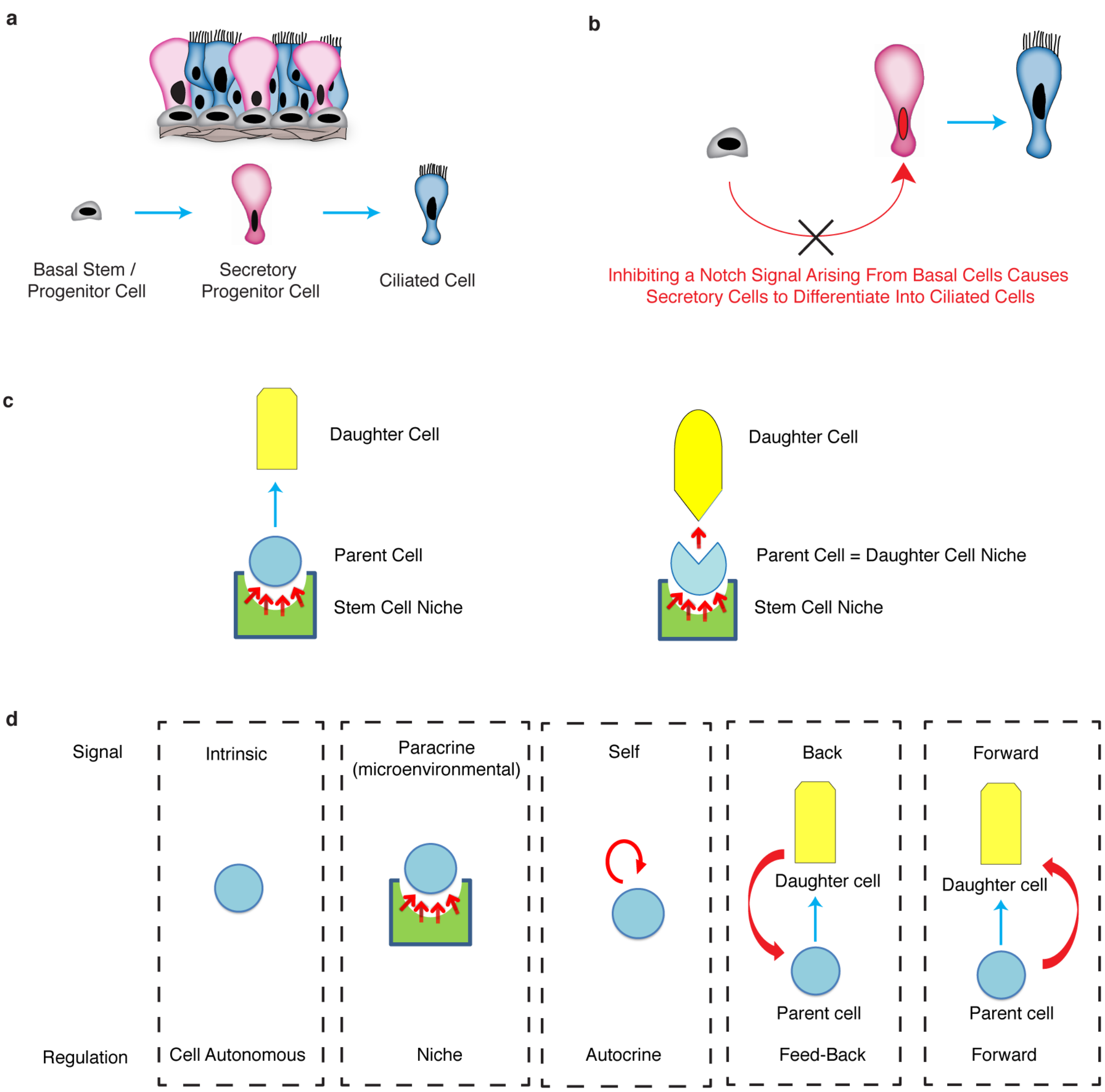
RNA extraction and quantitative RT–PCR. Total RNA was extracted from sorted airway epithelial cells from individual mice to analyse gene expression by quantitative RT–PCR. These procedures were performed as previously described[37]. Relative mRNA expression was normalized to baseline transcript levels in secretory progenitor cells in Figs 2a and 4a, and in control YFP^{+} cells in Fig. 3d, h. In addition, the primer sequences for the following genes were used: *Notch1*: forward 5′-tgagactgccaaagtgttgc-3′ and reverse 5′-gtgggagacagagtgggtgt-3′; *Notch2*: forward 5′-cctgaacgggcagtacattt-3′ and reverse 5′-gcgtagcccttcagacactc-3′; *Notch3*: forward 5′-tgagtgtccagctggctatg-3′ and reverse 5′-cacaggtgccattgtgtagg-3′; *Dll1*: forward 5′-ttagcatcattggggctacc-3′ and reverse 5′-taagtgttggggcgatcttc-3′; *Jag1*: forward 5′-cagtgcctctgtgagaccaa-3′ and reverse 5′-aggggtcagagagacaagca-3′; *Jag2*: forward 5′-cagatccgagtacgctgtga-3′ and reverse 5′-ggcttctttgcattctttgc-3′; *Hes1*: forward 5′-ctaccccagccagtgtcaac-3′ and reverse 5′-atgccgggagctatctttct-3′; *Hey1*: forward 5′-gagaccatcgaggtggaaaa-3′ and reverse 5′-agcagatccctgcttctcaa-3′; *HeyL*: forward 5′-ccccttaccctatctcagc-3′ and reverse 5′-acatggtgggattgggacta-3′; *RBPjk* exons 6–7: forward 5′-ggcagtggttggaagaaaaa-3′ and reverse 5′-atgtcatcgctgttgccata-3′; *Notch2* exon3: forward 5′-aacatcgagaccctgtgag-3′ and reverse 5′-ggctgagcatgtgacaggta-3′; *Jag2* exon2: Forward 5′-cgtgtgccttaaggagtacca-3′ and reverse 5′-gcgaactgaaagggaatgac-3′; *Scgb3a2*: forward 5′gacaggactgaagaagtgtgtgg-3′ and reverse 5′-ggaggttgttcacgtagcaaagg-3′; *c-myb*: forward 5′-gctgaagaagctggtggaac-3′ and reverse 5′-caacgcttcggaccatattt-3′.

Cell culture and viral transduction. Mouse tracheal epithelial cells were dissociated with papain and sorted with EpCAM and GSIβ4 as previously described[21,37]. Cells were cultured and expanded in complete SAGM (small airway epithelial cell growth medium; Lonza, CC-3118) using 5 mM Rock inhibitor Y-27632 (Selleckbio, S1049). To initiate air–liquid interface (ALI) cultures, airway basal cells were dissociated and seeded onto transwell membranes. After confluence, media was removed from the upper chamber. Mucocilary differentiation was performed with PneumaCult-ALI Medium (StemCell, 05001). Differentiation of airway basal cells on ALI was followed by directly visualizing beating cilia in real time after 8 days. One day after plating, mouse basal cells were infected with lentiviral vectors carrying shRNAs targeting mouse *Jag2*. Four different clones were obtained from Sigma (MISSION shRNA jagged2 NM_010588, clones TRCN0000028858, TRCN0000028871, TRCN0000028877, TRCN0000028906), and cloned into pLKO.1 vector (Addgene Plasmid 10878). Lentiviral production was performed in HEK293 cells following standard protocols. Concentrated viruses were used at a MOI of 6 to infect murine basal cells for 9 h at 37 °C in 5% CO_2, one day after plating. The cells were allowed to grow to confluence before being transferred onto transwell membranes. Then 23 days after ALI initiation, cells were washed, collected and sorted for GFP and cell specific markers. To assess the efficiency of shRNA Jag2 knockdown, non-purified infected cells were collected 72 h after infection and lysed in TRI Reagent.

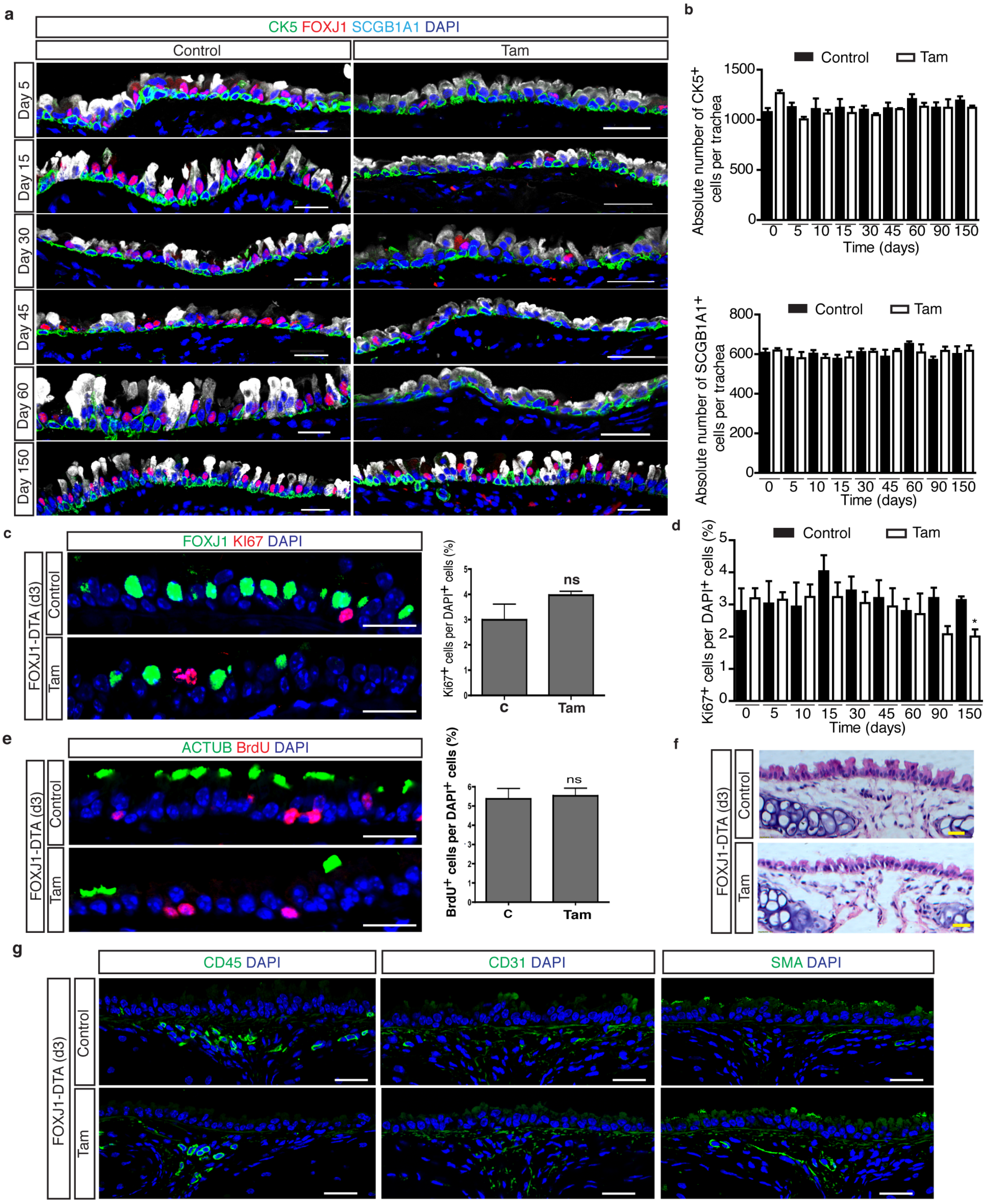
Statistical analysis. The standard error of the mean was calculated from the average of the indicated number of samples in each case ($n =$ biological replicates/condition/experiment). All the experiments were repeated at least three times with the exception of *CK5-rtTA;tet(O)Cre;Mindbomb1*$^{fl/fl}$ and the cell ablation experiments that were repeated twice. Data was compared among groups using the Student's *t*-test (unpaired, two-tailed test). A *P* value of less than 0.05 was considered significant. The analysis was performed with Prism software (Graphpad Prism version 5.0a).

Data reporting. No statistical methods were used to predetermine sample size. Experiments were performed completely blinded, being repeated by two different investigators and some of the experiments were repeated without knowing which samples were analysed.

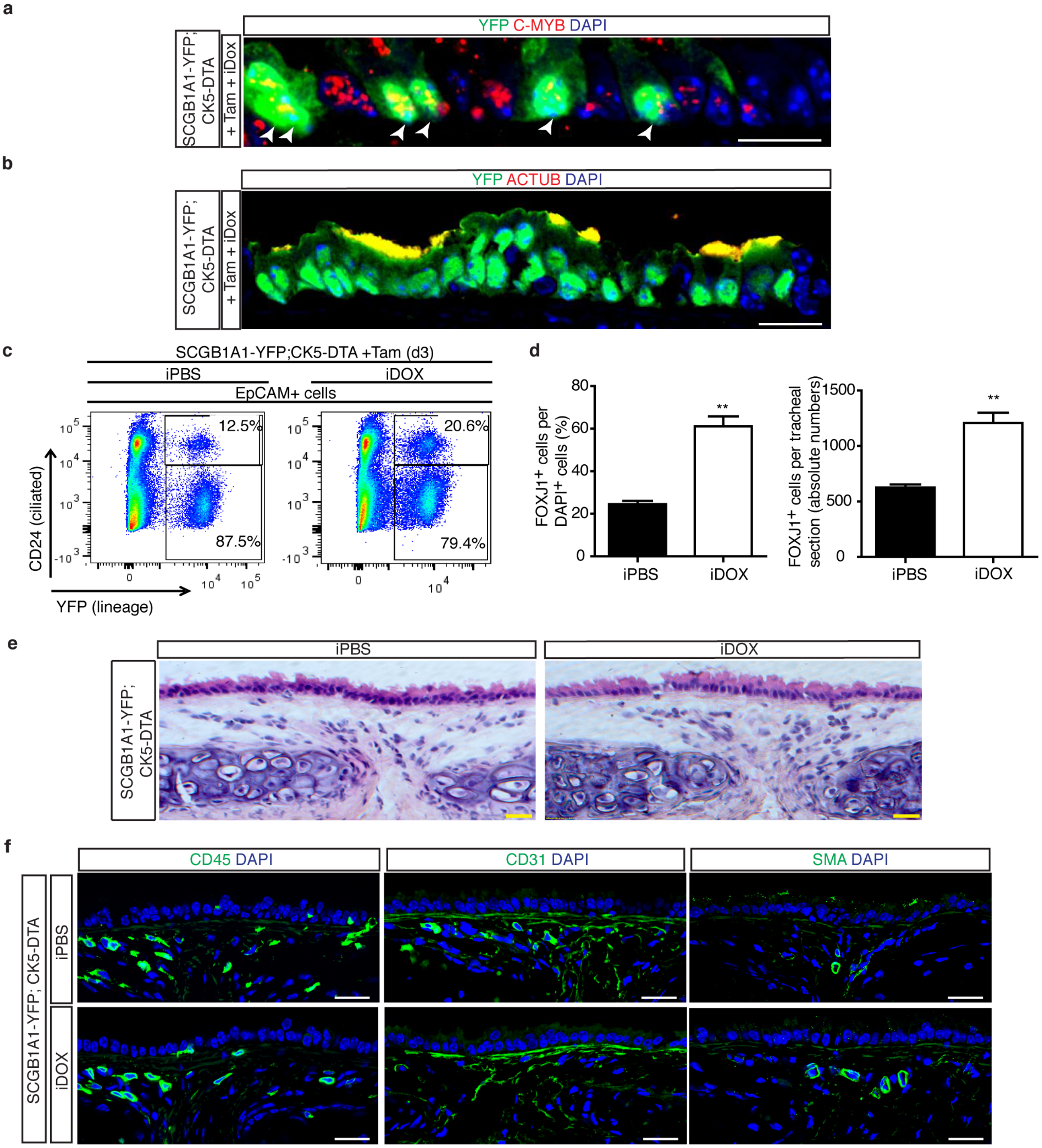
31. Diamond, I., Owolabi, T. & Marco, M. Conditional gene expression in the epidermis of transgenic mice using the tetracycline-regulated transactivators tTA and rTA linked to the keratin 5 promoter. *J. Invest. Dermatol.* **115,** 788–794 (2000).
32. Van Keymeulen, A. *et al.* Distinct stem cells contribute to mammary gland development and maintenance. *Nature* **479,** 189–193 (2011).
33. Weber, T. *et al.* Inducible gene expression in GFAP$^+$ progenitor cells of the SGZ and the dorsal wall of the SVZ-A novel tool to manipulate and trace adult neurogenesis. *Glia* **59,** 615–626 (2011).
34. Tanigaki, K. *et al.* Notch-RBP-J signaling is involved in cell fate determination of marginal zone B cells. *Nature Immunol.* **3,** 443–450 (2002).
35. Xu, J., Krebs, L. T. & Gridley, T. Generation of mice with a conditional null allele of the Jagged2 gene. *Genesis* **48,** 390–393 (2010).
36. Kim, J. K. *et al. In vivo* imaging of tracheal epithelial cells in mice during airway regeneration. *Am. J. Respir. Cell Mol. Biol.* **47,** 864–868 (2012).
37. Pardo-Saganta, A., Law, B. M., Gonzalez-Celeiro, M., Vinarsky, V. & Rajagopal, J. Ciliated cells of pseudostratified airway epithelium do not become mucous cells after ovalbumin challenge. *Am. J. Respir. Cell Mol. Biol.* **48,** 364–373 (2013).


a
Basal Stem / Progenitor Cell
Secretory Progenitor Cell
Ciliated Cell
b
Inhibiting a Notch Signal Arising From Basal Cells Causes Secretory Cells to Differentiate Into Ciliated Cells
c
Daughter Cell
Parent Cell
Stem Cell Niche
Daughter Cell
Parent Cell = Daughter Cell Niche
Stem Cell Niche
d
Signal
Intrinsic
Paracrine (microenvironmental)
Self
Back
Forward
Daughter cell
Parent cell
Daughter cell
Parent cell
Regulation
Cell Autonomous
Niche
Autocrine
Feed-Back
Forward

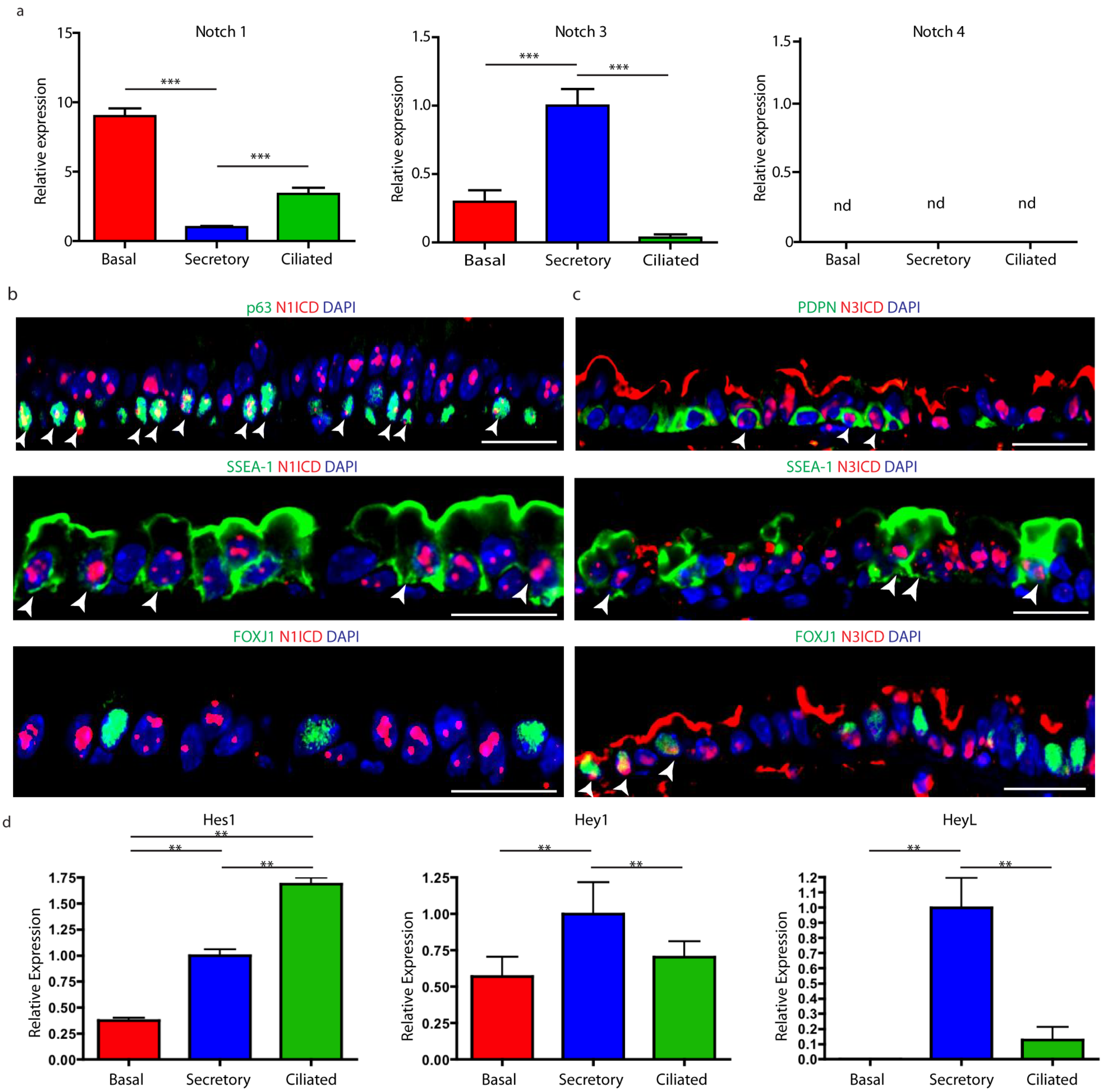
Extended Data Figure 1 | Parent stem/progenitor cells can serve as niches for their own daughter cells. **a**, Schematic representation of the airway epithelial cell lineage. Basal stem/progenitor cells give rise to secretory progenitor cells that, in turn, give rise to terminally differentiated ciliated cells. **b**, Basal cells expressing Notch ligands provide a tonic forward Notch signal to neighbouring secretory daughter cells. Blocking this forward signal prevents Notch activation in secretory cells and results in their differentiation into ciliated cells. **c**, A schematic of the traditional arrangement of a stem cell that is maintained in a stem cell niche (left) and a schematic that further illustrates that stem cells can themselves serve as daughter cell niches, in which the parent stem cell itself is required for the maintenance of its own progeny (right). **d**, Schematic of the types of signals that occur between cells within a lineage and the theoretical modes of cell regulation that they imply. Blue arrows indicate a lineage relationship. Red arrows represent signals.



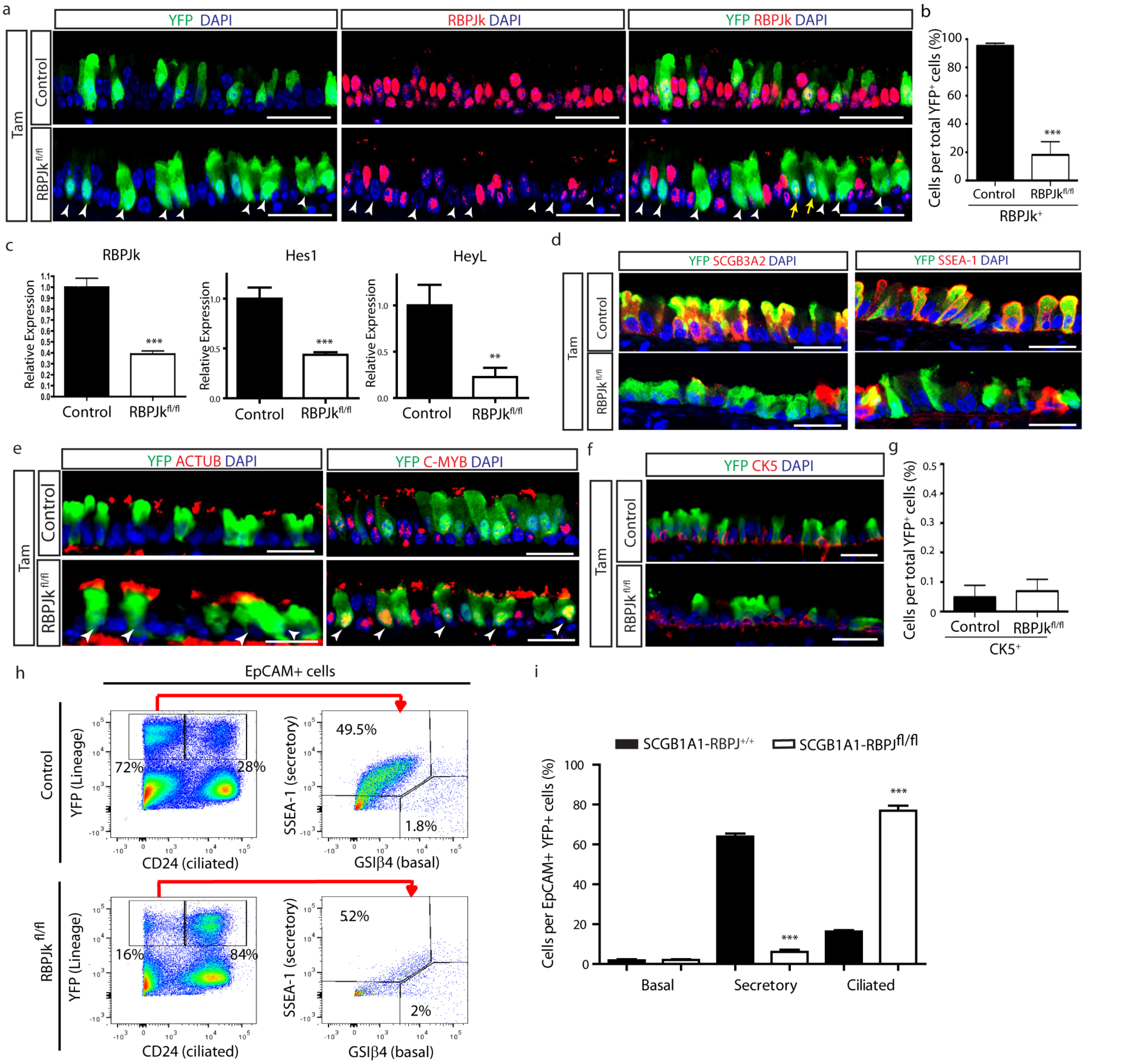
Extended Data Figure 2 | Ablation of ciliated cells has no effect on airway cell proliferation, mesenchymal cell types, mesenchymal morphology and airway stem and progenitor cell replication over time. **a**, Immunostaining for basal (CK5 (green)), ciliated (FOXJ1 (red)), and secretory cells (SCGB1A1 (white)) on either control (left panels) or tamoxifen (Tam)-treated FOXJ1-DTA mice (right panels) 5, 15, 30, 45, 60 and 150 days after ciliated cell ablation ($n = 3$ mice). **b**, Quantification of absolute cell numbers of basal CK5$^+$ cells (top graph) and secretory SCGB1A1$^+$ cells (bottom graph) per trachea on control (black bars) or Tam-treated (white bars) mice over time ($n = 3$ mice). **c**, Immunostaining for ciliated cells (FOXJ1 (green)) and proliferating cells (Ki67 (red)) on either control (upper panel) or tamoxifen (Tam)-treated FOXJ1-DTA mice (lower panel) ($n = 6$ mice). On the right, quantification of the percentage of Ki67$^+$ cells per total DAPI$^+$ cells in tracheal sections from control (C) or Tam-treated mice 3 days after cell ablation ($n = 3$ mice). **d**, Quantification of the percentage of proliferating Ki67$^+$ cells relative to total DAPI$^+$ cells in control (black bars) or Tam-treated (white bars) mice over time ($n = 3$ mice). **e**, Immunostaining for ciliated cells (AcTub (green)) and cells that have undergone proliferation (BrdU (red)) on either control (upper panel) or Tam-treated mice (lower panel) at day 3 ($n = 6$ mice). On the right, quantification of the percentage of BrdU$^+$ cells per total DAPI$^+$ cells in tracheal sections from control (C) or Tam treated mice ($n = 3$ mice). **f**, Haematoxylin & eosin (H&E) staining of tracheal sections 3 days after ciliated cell ablation. **g**, Immunostaining for CD45$^+$ haematopoietic cells (left panels), CD31$^+$ endothelial cells (middle panels) and SMA$^+$ smooth muscle cells (right panels) (green) three days after cell ablation ($n = 6$ mice). Nuclei stained with DAPI (blue). The ns indicates that the cell number comparisons are not statistically significant. n = biological replicates/condition (two independent experiments). Data shown in the graphs are means ± s.e.m. Scale bar, 20 µm.



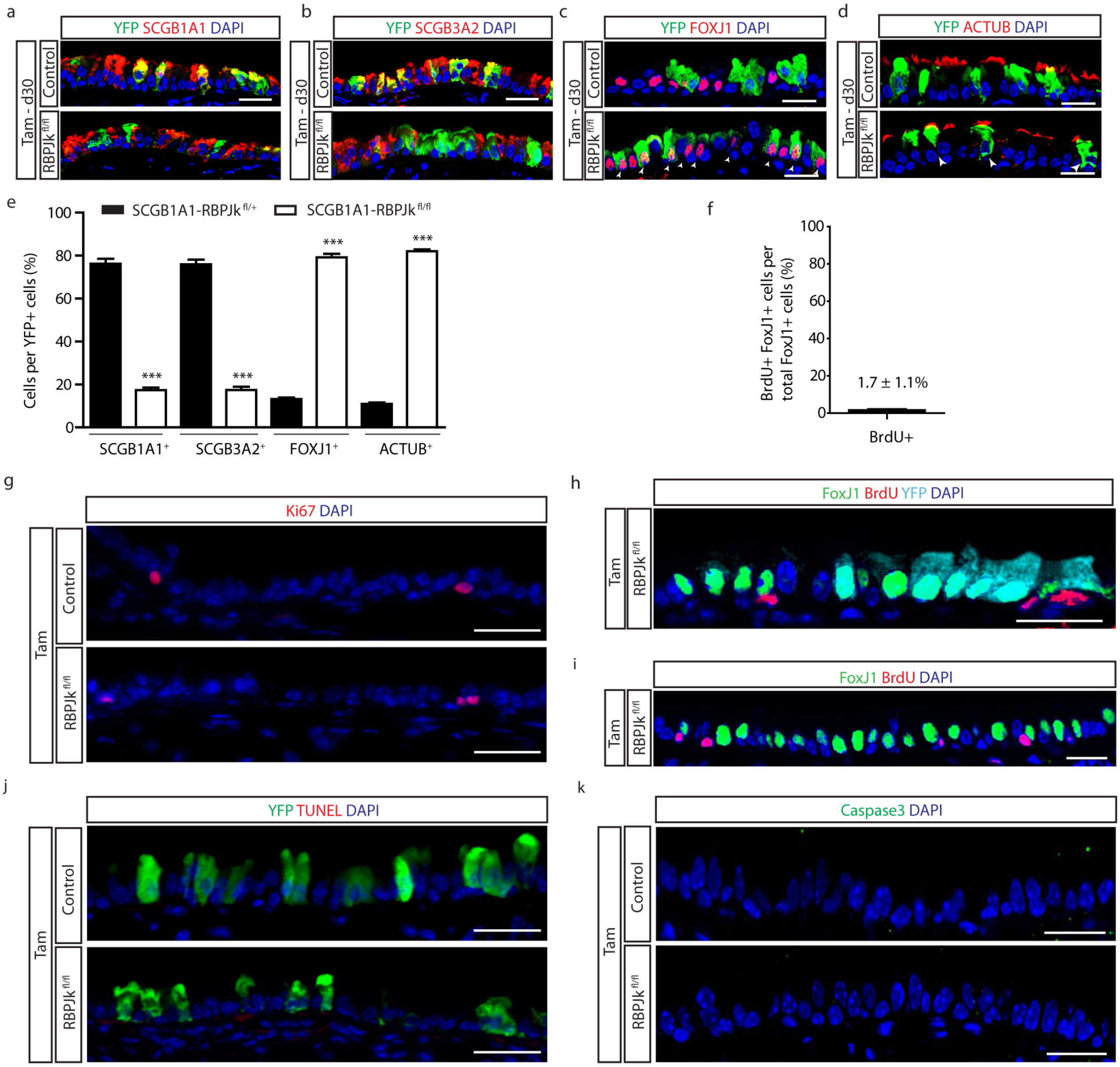
Extended Data Figure 3 | Basal stem/progenitor cell ablation promotes the differentiation of secretory cells into ciliated cells without affecting the mesenchyme. **a**, Immunostaining for YFP lineage label (green) and the ciliated cell marker c-MYB (red) in SCGB1A1-YFP; CK5-DTA mice (*n* = 3 mice). White arrowheads point to double positive cells. **b**, Immunostaining for YFP lineage label (green) and the ciliated cell marker AcTub (red) using SCGB1A1-YFP; CK5-DTA mice (*n* = 3 mice). **c**, Flow cytometry analysis for lineage-labelled YFP$^+$ cells (*x* axis) and CD24$^+$ ciliated cells (*y* axis) cells from control iPBS-treated or doxycycline-treated SCG1A1-YFP; CK5-DTA mice. **d**, Quantification of the percentage of FOXJ1$^+$ cells per total DAPI$^+$ cells in tracheal sections from control iPBS-treated or iDOX-treated SCG1A1-YFP; CK5-DTA mice (*n* = 3 mice). On the right, absolute numbers of FOXJ1$^+$ cells per tracheal section (*n* = 3 mice). **e**, H&E staining of tracheal sections following basal cell ablation. **f**, Immunostaining for CD45$^+$ haematopoietic cells (left panels), CD31$^+$ endothelial cells (middle panels) and SMA$^+$ smooth muscle cells (right panels) (green) in control or basal cell-ablated trachea (*n* = 3 mice). All analyses were performed 3 days after cell ablation. Nuclei stained with DAPI (blue). *n* = biological replicates/condition (two independent experiments). $^{**}P < 0.01$. Data shown in the graphs are means ± s.e.m. Scale bar, 20 μm.



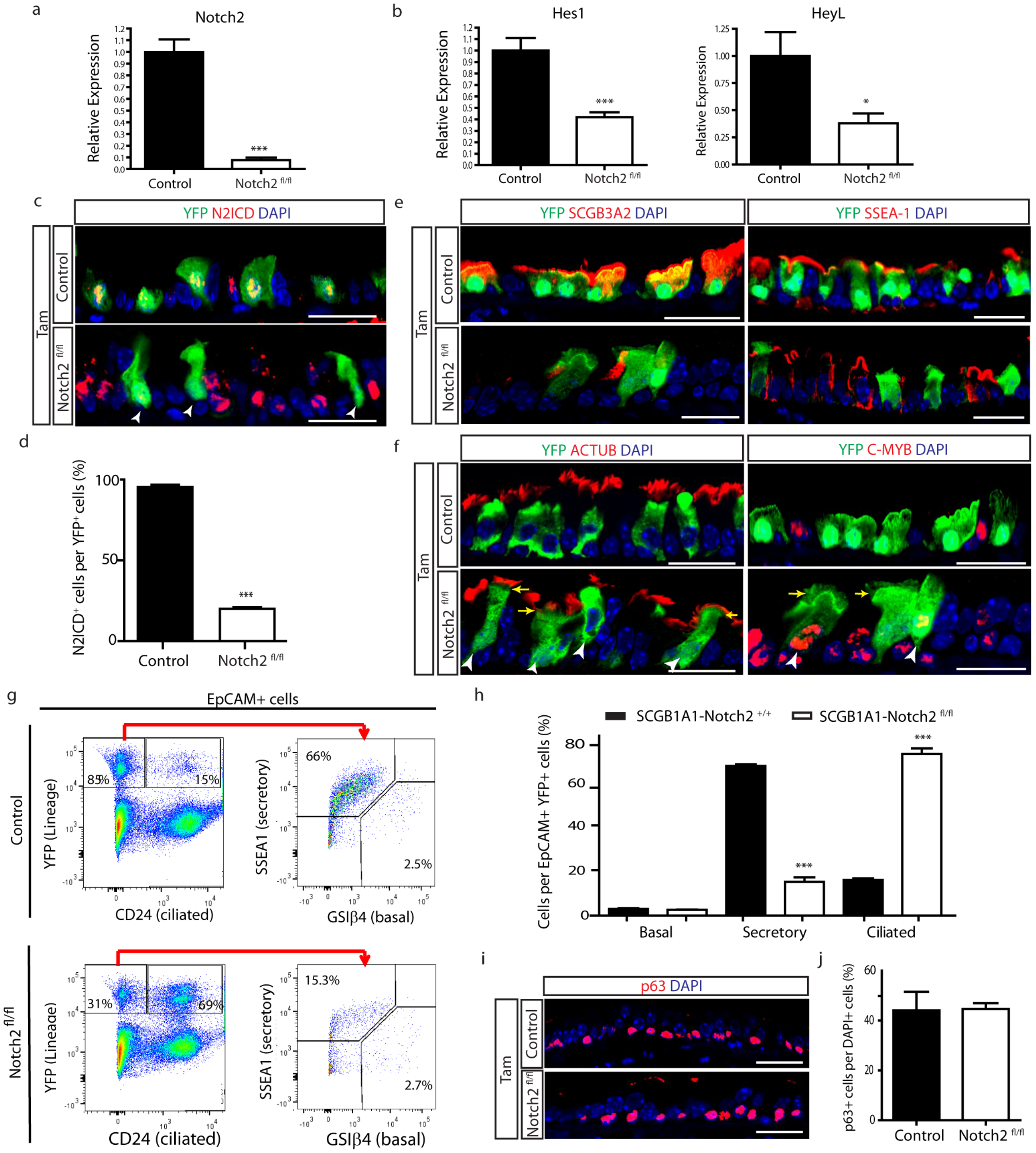
Extended Data Figure 4 | Characterization of Notch pathway components in the steady-state murine tracheal epithelium. a, Relative mRNA expression of *Notch1*, *Notch3* and *Notch4* assessed by qRT–PCR in pure sorted populations of airway epithelial cells ($n = 3$ mice). Relative expression is normalized to baseline transcript levels in secretory progenitor cells. **b**, Immunostaining for N1ICD (red) in combination with the basal cell marker p63 (top panel), the secretory cell marker SSEA-1 (middle panel) and the ciliated cell marker FOXJ1 (bottom panel) (green). **c**, Immunostaining for N3ICD (red) in combination with the basal cell marker podoplanin (PDPN) (top panel), the secretory cell marker SSEA-1 (middle panel) and the ciliated cell marker FOXJ1 (bottom panel) (green). **d**, Relative mRNA expression of *Hes1*, *Hey1* and *HeyL* assessed by qRT–PCR in pure sorted populations of airway epithelial cells ($n = 3$ mice). Relative expression is normalized to baseline transcript levels in secretory progenitor cells. n = biological replicates/condition. $^{**}P < 0.01$; $^{***}P < 0.001$. nd indicates lack of detection. Data shown in the graphs are means ± s.e.m. Nuclei stained with DAPI (blue). White arrowheads point to double positive cells. Scale bar, 20 μm.



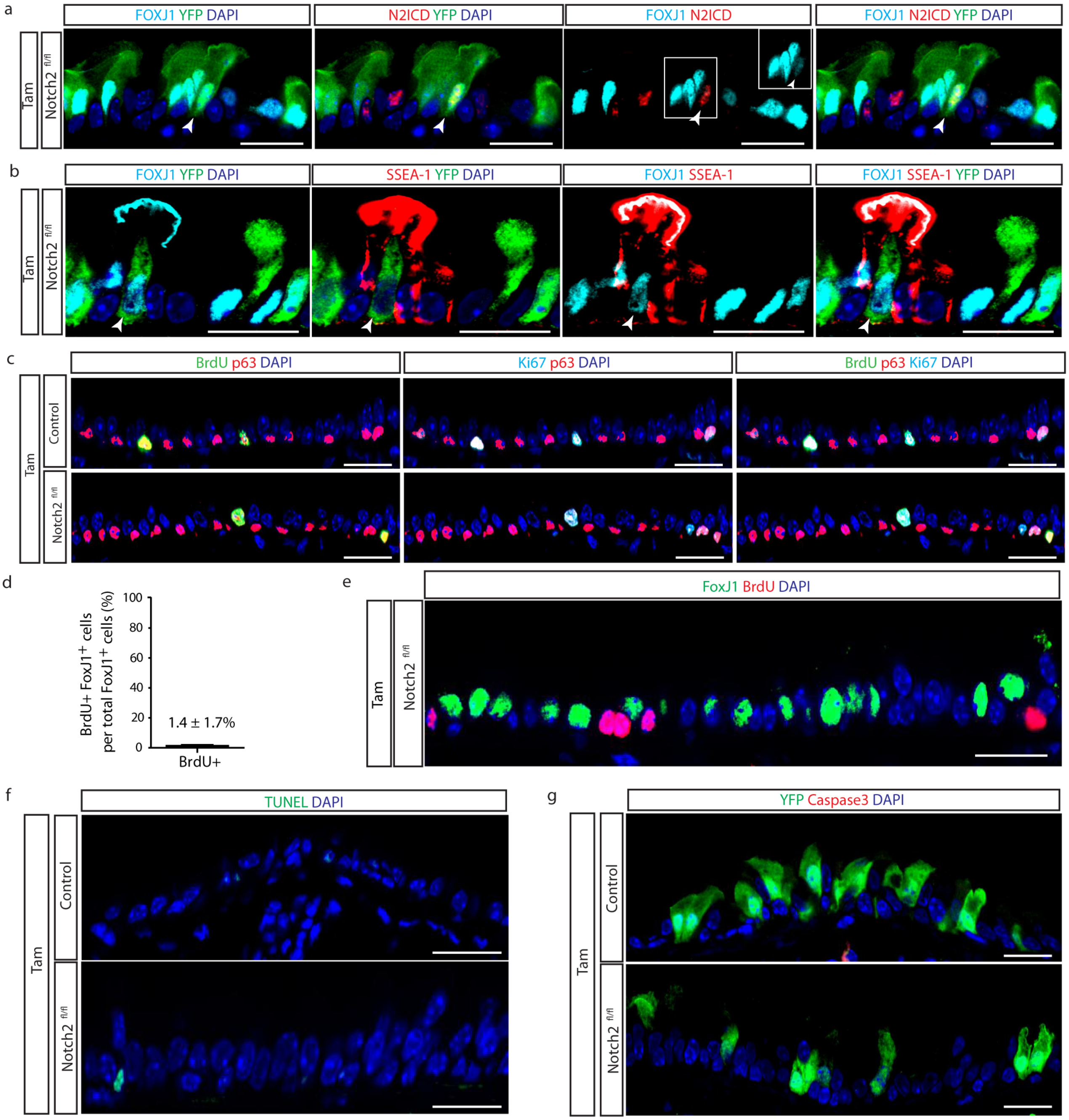
Extended Data Figure 5 | Downregulation of Notch signalling transduction following *RBPjk* deletion in secretory progenitor cells induces their conversion into ciliated cells. **a**, Immunostaining for lineage-labelled YFP$^+$ cells (green) in combination with RBPjk (red) in Tam-treated SCGB1A1-RBPjk$^{fl/+}$ control mice (upper panels) and Tam-treated SCGB1A1-RBPjk$^{fl/fl}$ mice (lower panels). White arrowheads point to lineage-labelled RBPjk$^-$ cells. The yellow arrows point to lineage-labelled cells that have not undergone recombination. **b**, Quantification of the percentage of RBPjk$^+$ cells per total YFP$^+$ cells at experimental day 15 following tamoxifen administration to SCGB1A1-RBPjk$^{fl/+}$ control (black bar) and SCGB1A1-RBPjk$^{fl/fl}$ mice (white bar) ($n = 6$ mice). **c**, Relative mRNA expression of Notch signalling component genes (*RBPjk*, *Hes1*, *HeyL*) analysed by qRT–PCR in sorted YFP$^+$ cells from Tam-treated SCGB1A1-RBPjk$^{+/+}$ control mice (black bars) ($n = 3$ mice) and Tam-treated SCGB1A1-RBPjk$^{fl/fl}$ mice (white bars) ($n = 4$ mice). Relative expression is normalized to baseline transcript levels in YFP$^+$ control cells. **d**, Immunostaining for YFP lineage label (green) and the secretory progenitor cell markers SCGB3A2 (left panels) and SSEA-1 (right panels) (red) in Tam-treated SCGB1A1-RBPjk$^{fl/+}$ mice (control) (top panels) and SCGB1A1-RBPjk$^{fl/fl}$ mice (bottom panels). **e**, Immunostaining for YFP lineage label (green) and the ciliated cell markers AcTub (left panels) and c-MYB (right panels) (red) in Tam-treated SCGB1A1-RBPjk$^{fl/+}$ mice (control) (top panels) and SCGB1A1-RBPjk$^{fl/fl}$ mice (bottom panels). White arrowheads point to lineage-labelled secretory cells that differentiated into ciliated cells following *RBPjk* deletion. **f**, Immunostaining for lineage-labelled YFP$^+$ cells (green) and the basal cell marker CK5 (red) on either Tam-treated SCGB1A1-RBPjk$^{fl/+}$ control mice (upper panel) or Tam-treated SCGB1A1-RBPjk$^{fl/fl}$ mice (lower panel). **g**, Quantification of the percentage of CK5$^+$ cells per total YFP$^+$ cells in Tam-treated SCGB1A1-RBPjk$^{fl/fl}$ mice compared to control mice. **h**, Flow cytometry analysis of EpCAM$^+$ YFP$^+$ CD24$^+$ lineage-labelled ciliated cells and EpCAM$^+$YFP$^+$CD24$^-$ SSEA-1$^+$ lineage-labelled secretory cells or EpCAM$^+$YFP$^+$CD24$^-$GSIβ4$^+$ lineage-labelled basal cells in airways from either control or Tam-treated SCGB1A1-RBPjk$^{fl/fl}$ mice. **i**, Quantification of the percentage of epithelial (EpCAM$^+$) lineage-labelled (YFP$^+$) basal, secretory and ciliated cells in either Tam-treated SCGB1A1-RBPjk$^{+/+}$ control or SCGB1A1-RBPjk$^{fl/fl}$ mice by flow cytometry ($n = 3$ mice). The analysis was performed 10 days after the last tamoxifen injection. Images are representative of $n = 6$ mice per condition (biological replicates) repeated three times. Nuclei stained with DAPI (blue). $**P < 0.01$; $***P < 0.001$. Data shown in the graphs are means ± s.e.m. Scale bar, 20 µm.



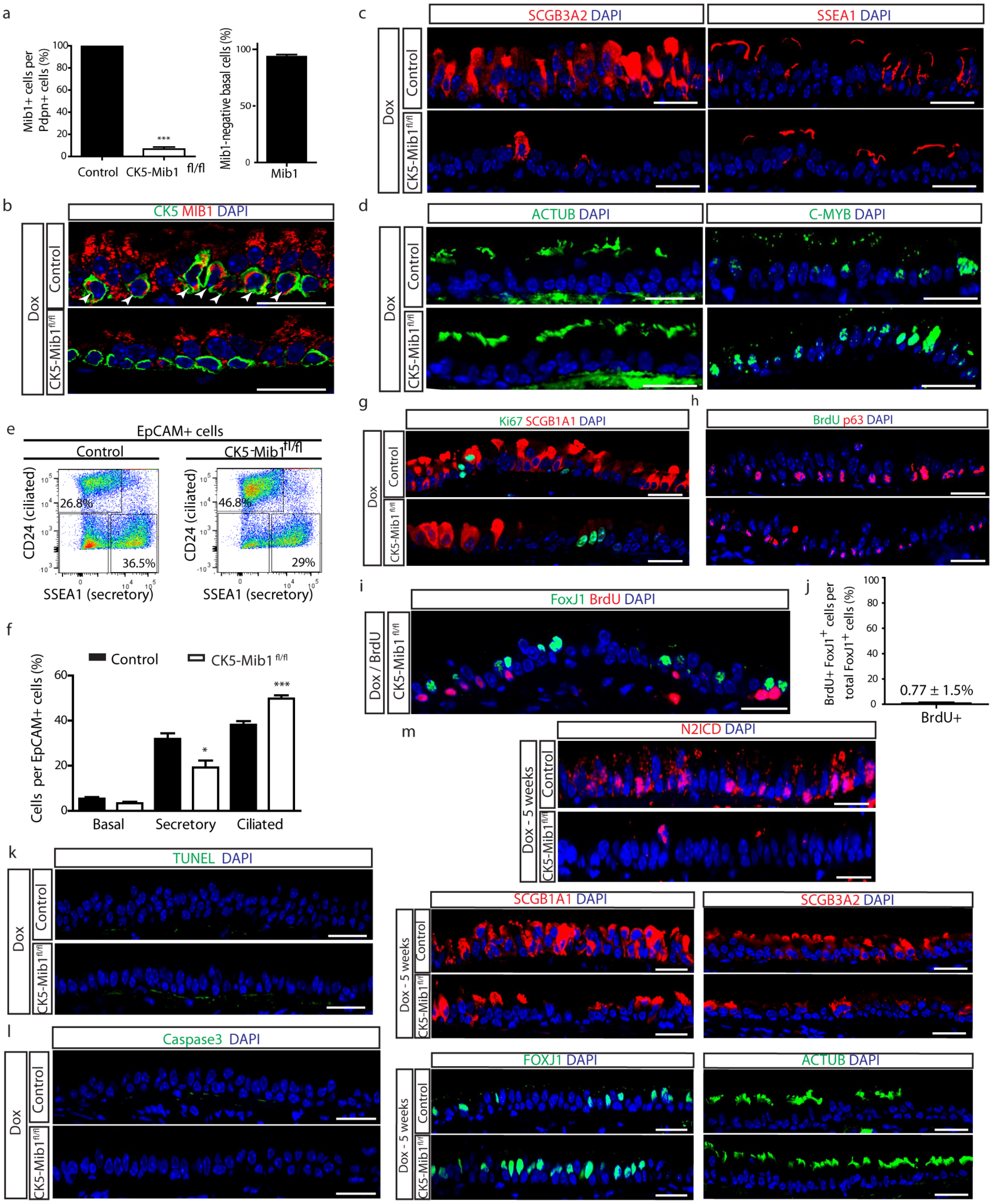
Extended Data Figure 6 | Lineage-labelled ciliated cells demonstrate long term persistence after *RBPjk* deletion without a change in epithelial cell proliferation and apoptosis. **a**–**d**, Immunostaining for the lineage label YFP (green) in combination with the secretory cell markers SCGB1A1 (**a**), SCGB3A2 (**b**) or the ciliated cell markers FOXJ1 (**c**) and AcTub (**d**) (red) on either Tam-treated SCGB1A1-RBPjk$^{fl/+}$ control mice (upper panels) or Tam-treated SCGB1A1-RBPjk$^{fl/fl}$ mice (lower panels) 30 days after the last tamoxifen injection ($n = 3$ mice). White arrowheads point to lineage-labelled ciliated cells. **e**, Quantification of the percentage of each cell type per YFP^{+} cells on either control mice (black bars) or Tam-treated SCGB1A1-RBPjk$^{fl/fl}$ mice (white bars) at day 30. **f**, Quantification of the percentage of ciliated FOXJ1^{+} cells that incorporate BrdU after continuous BrdU administration to Tam-treated SCGB1A1-RBPjk$^{fl/fl}$ mice ($n = 3$ mice). **g**, Immunostaining for Ki67 (red) to assess overall proliferation in either Tam-treated SCGB1A1-RBPjk$^{fl/+}$ control mice (upper panel) or Tam-treated SCGB1A1-RBPjk$^{fl/fl}$ mice (lower panel) ($n = 3$ mice). **h**, **i**, Immunostaining for FOXJ1 (green) and BrdU (red) in combination with YFP (cyan) (**h**) or alone (**i**) on Tam-treated SCGB1A1-RBPjk$^{fl/fl}$ mice that received continuous BrdU ($n = 3$ mice). **j**, Immunostaining to detect apoptotic cells by TUNEL assay (red) in combination with YFP lineage-labelled cells (green) in either Tam-treated SCGB1A1-RBPjk$^{fl/+}$ control mice (upper panel) or Tam-treated SCGB1A1-RBPjk$^{fl/fl}$ mice (lower panel) ($n = 3$ mice). **k**, Immunostaining for activated caspase3 (green) in control and Tam-treated SCGB1A1-RBPjk$^{fl/fl}$ mice ($n = 3$ mice). **f**–**k**, Analysis conducted 10 days after induction. Nuclei stained with DAPI (blue). n = biological replicates per condition. ***$P < 0.001$. Data shown in the graph are means ± s.e.m. Scale bar, 20 μm.



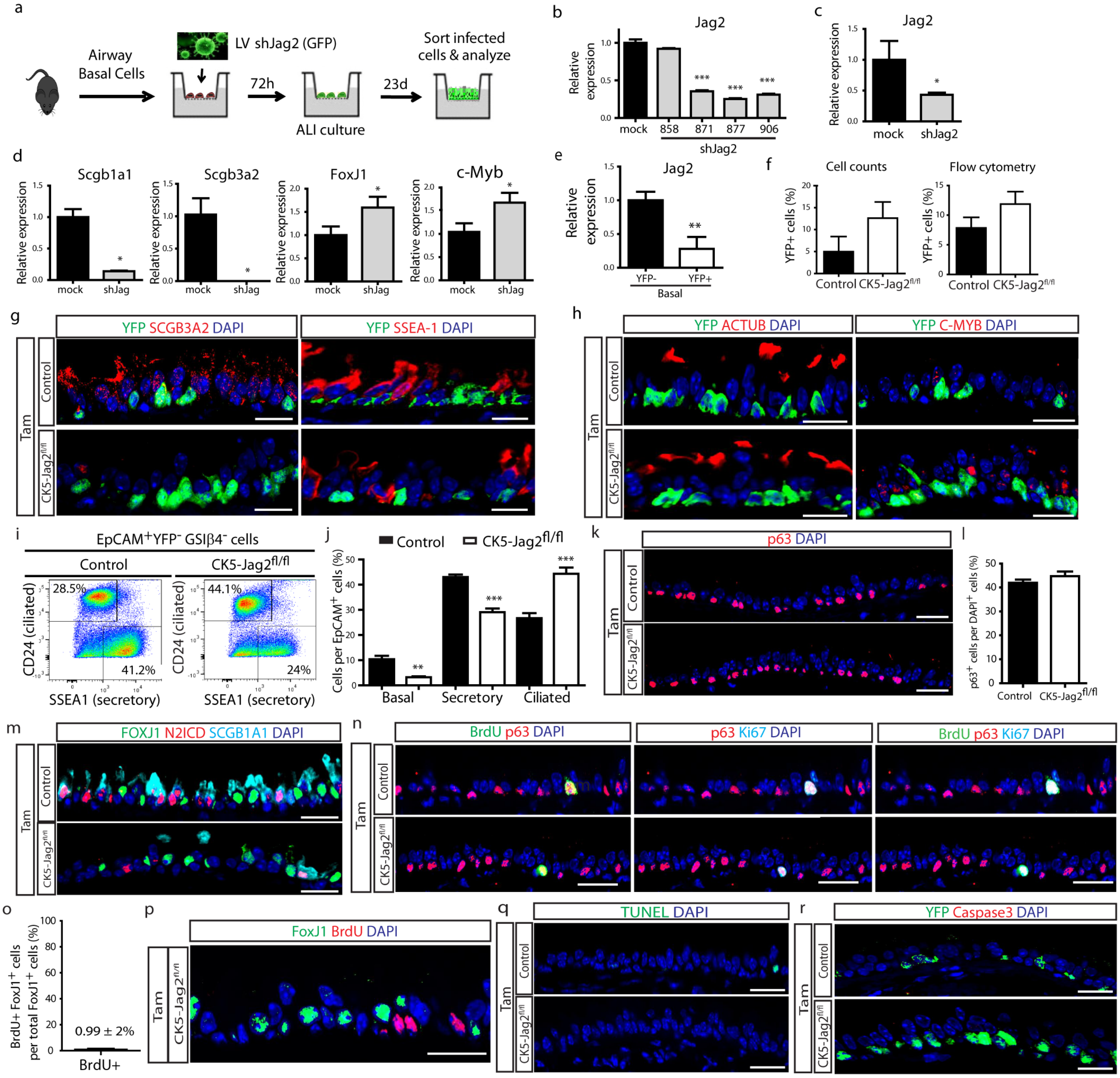
Extended Data Figure 7 | Efficient deletion of *Notch2* in secretory progenitor cells and its effect on cell type distribution. **a**, Relative mRNA expression of *Notch2* in YFP$^+$ cells from Tam-treated SCGB1A1-Notch2$^{+/+}$ control mice and Tam-treated SCGB1A1-Notch2$^{fl/fl}$ experimental mice assessed by qRT–PCR ($n = 3$ mice). **b**, Relative mRNA expression of the Notch target genes (*Hes1, HeyL*) in YFP$^+$ cells from control mice and Tam-treated SCGB1A1-Notch2$^{fl/fl}$ experimental mice ($n = 3$ mice). Relative expression is normalized to baseline transcript levels in lineage-labelled YFP$^+$ control cells. **c**, Immunostaining for lineage label YFP (green) in combination with N2ICD (red) on control mice (Tam-treated SCGB1A1-Notch2$^{+/+}$) and experimental airways (Tam-treated SCGB1A1-Notch2 $^{fl/fl}$). White arrowheads point to lineage-labelled cells that had lost Notch2 and therefore do not show N2ICD expression. **d**, Quantification of the percentage of N2ICD$^+$ cells per total YFP$^+$ cells in Tam-treated SCGB1A1-Notch2$^{fl/fl}$ mice compared to control ($n = 7$ mice). **e**, Immunostaining for YFP lineage label (green) and the secretory progenitor cell markers SCGB3A2 (left panels) and SSEA-1 (right panels) (red) in control (top panels) and experimental (bottom panels) mice. **f**, Immunostaining for YFP lineage label (green) and the ciliated cell markers AcTub (left panels) and c-MYB (right panels) (red) in control (top panels) and experimental (bottom panels) mice. White arrowheads point to lineage-labelled secretory cells that differentiated into ciliated cells following *Notch2* deletion. Yellow arrows point to actual cilia (green) in lineage-labelled cells. **g**, Flow cytometry analysis of EpCAM$^+$YFP$^+$CD24$^+$ lineage-labelled ciliated cells and EpCAM$^+$ YFP$^+$CD24$^-$SSEA-1$^+$ lineage-labelled secretory cells or EpCAM$^+$YFP$^+$CD24$^-$GSIβ4$^+$ lineage-labelled basal cells in airways from either Tam-treated SCGB1A1-Notch2$^{+/+}$ control mice or Tam-treated SCGB1A1-Notch2$^{fl/fl}$ mice. **h**, Quantification of the percentage of epithelial (EpCAM$^+$) lineage-labelled (YFP$^+$) basal, secretory and ciliated cells in either Tam-treated SCGB1A1-Notch2$^{+/+}$ control ($n = 4$ mice) or SCGB1A1-Notch2$^{fl/fl}$ mice ($n = 6$ mice) by flow cytometry. **i**, Immunostaining for the basal cell transcription factor p63 (red) on control or SCGB1A1-Notch2$^{fl/fl}$ airways. **j**, Quantification of the percentage of p63$^+$ cells per total DAPI$^+$ cells on tracheal sections from control or experimental mice ($n = 7$ mice). Analysis performed 10 days after induction. Images are representative of $n = 7$ mice per condition (biological replicates) repeated three times (three independent experiments). Nuclei stained with DAPI (blue). $^*P < 0.05$; $^{***}P < 0.001$. Data shown in the graphs are means ± s.e.m. Scale bar, 20 μm.



Extended Data Figure 8 | Proliferation and apoptosis following deletion of *Notch2* in secretory progenitor cells. **a**, Immunostaining for lineage label YFP (green), FOXJ1 (cyan) and N2ICD (red) in Tam-treated SCGB1A1-Notch2$^{fl/fl}$ mice. White arrowhead points to a lineage-labelled cell co-expressing markers for secretory and ciliated cell fates. The inset shows the single stain for FOXJ1 of the indicated region. **b**, Immunostaining for lineage label YFP (green), FOXJ1 (cyan) and SSEA-1 (red) in Tam-treated SCGB1A1-Notch2$^{fl/fl}$ mice. White arrowhead points to a lineage-labelled transitional cell. **c**, Immunostaining for BrdU (green), p63 (red) and Ki67 (cyan) to assess overall proliferation on either Tam-treated SCGB1A1-Notch2$^{+/+}$ control mice (upper panels) or Tam-treated SCGB1A1-Notch2$^{fl/fl}$ mice (lower panels). **d**, Quantification of the percentage of ciliated FOXJ1^{+} cells that incorporate BrdU after continuous BrdU administration to Tam-treated SCGB1A1-Notch2$^{fl/fl}$ mice ($n = 4$ mice). **e**, Immunostaining for FOXJ1 (green) and BrdU (red) on Tam-treated SCGB1A1-Notch2$^{fl/fl}$ mice that received continuous BrdU ($n = 4$ mice). **f**, Immunostaining to detect apoptotic cells by TUNEL assay (green) on either Tam-treated SCGB1A1-Notch2$^{+/+}$ control mice (upper panel) or Tam-treated SCGB1A1-Notch2$^{fl/fl}$ mice (lower panel). **g**, Immunostaining for YFP (green) in combination with activated caspase3 (red) on control mice (upper panel) or Tam-treated SCGB1A1-Notch2$^{fl/fl}$ mice (lower panel). Analysis performed 10 days after induction. Images are representative of $n = 7$ mice per condition (biological replicates) repeated three times (three independent experiments). Nuclei stained with DAPI (blue). Scale bar, 20 μm.



Extended Data Figure 9 | Loss of Notch ligands in basal stem cells promotes secretory cell differentiation into ciliated cells without affecting proliferation or apoptosis. **a**, Quantification of the percentage of basal PDPN$^+$ cells that express Mib1 (left graph) on either Dox-treated CK5-Mib1$^{+/+}$ control mice or Dox-treated CK5-Mib1$^{fl/fl}$ mice ($n = 4$ mice). Right graph, percentage of basal cells in which Mib1 was deleted in Dox-treated CK5-Mib1$^{fl/fl}$ mice ($n = 4$ mice). **b**, Immunostaining for Mib1 (red) and the basal cell marker CK5 (green). White arrowheads point to Mib1$^+$ basal cells. **c**, Immunostaining for the secretory cell markers SCGB3A2 (left panels) and SSEA-1 (right panels) (red) in control (top panels) and experimental (bottom panels) mice. **d**, Immunostaining for the ciliated cell markers AcTub (left panels) and c-MYB (right panels) (green) in control (top panels) and experimental (bottom panels) mice. **e**, Flow cytometry analysis of EpCAM$^+$ CD24$^+$ ciliated cells and EpCAM$^+$ SSEA-1$^+$ secretory cells from control and experimental mice. **f**, Percentage of epithelial (EpCAM$^+$) basal, secretory and ciliated cells on both groups by flow cytometry ($n = 3$ mice). **g**, Immunostaining for Ki67 (green) and the secretory cell marker SCGB1A1 (red) on control (top panel) or Dox-treated CK5-Mib1$^{fl/fl}$ mice (bottom panel). **h**, Immunostaining for BrdU (green) in combination with the basal cell transcription factor p63 (red) on both groups. **i**, Immunostaining for FOXJ1 (green) and BrdU (red) on Dox-treated CK5-Mib1$^{fl/fl}$ mice that received continuous BrdU. **j**, Percentage of ciliated FOXJ1$^+$ cells that incorporate BrdU after continuous BrdU administration to Dox-treated CK5-Mib1$^{fl/fl}$ mice ($n = 4$ mice). **k**, Immunostaining to detect apoptotic cells by TUNEL assay (green) on either control (upper panel) or experimental mice (lower panel). **l**, Immunostaining for activated caspase3 (green) on both groups. **m**, Immunostaining for N2ICD (red), SCGB1A1 and SCGB3A2 (red), or FOXJ1 and AcTub (green) in control (top panels) or experimental mice (bottom panels) after five weeks of continuous doxycycline treatment ($n = 4$ mice). **a–l**, Analysis performed 2 weeks after the beginning of Dox induction. Images are representative of $n = 4$ mice per condition (biological replicates) repeated twice. *$P < 0.05$; ***$P < 0.001$. Data shown in the graphs are means ± s.e.m. Nuclei, DAPI (blue). Scale bar, 20 µm.



Extended Data Figure 10 | Disruption of *Jag2* in basal stem/progenitor cells causes the differentiation of secretory progenitor cells into ciliated cells without affecting proliferation or apoptosis. **a**, Schematic representation of *Jag2* inhibition using lentiviruses (LV) carrying shRNAs. Infected GFP$^+$ cells were cultured in ALI culture system for 23 days, when they were collected, sorted and analysed. **b**, Relative mRNA expression of *Jag2* in tracheal epithelial cells infected with mock vector (control) or with vectors carrying 4 different shRNAs targeting *Jag2* 72 h after infection. **c**, Relative mRNA expression of *Jag2* in tracheal epithelial basal cells infected with mock vector (control) or with lentivirus targeting *Jag2* (shJag2 877) after 23 days in ALI. **d**, Relative mRNA expression of the secretory genes (*Scgb1a1* and *Scgb3a2*) and the ciliated cell genes (*FoxJ1* and *c-myb*) in mock (black bars) and shJag2 877 (grey bars) infected cells 23 days after ALI initiation. Relative expression is normalized to baseline transcript levels in mock-infected cells. **e**, Relative mRNA expression of *Jag2* on sorted recombined (YFP$^+$) basal cells and unrecombined YFP$^-$ basal cells from Tam-treated CK5-Jag2$^{fl/fl}$ mice ($n = 3$ mice). Relative expression is normalized to baseline transcript levels in YFP$^-$ cells. **f**, Percentage of YFP$^+$ cells per total DAPI$^+$ cells (efficiency of recombination) on either Tam-treated CK5-Jag2$^{+/+}$ control (black bars) or Tam-treated CK5-Jag2$^{fl/fl}$ (white bars) mice assessed by manual counting (left graph) ($n = 5$ mice) or by flow cytometry (right graph) ($n = 3$ mice). **g**, Immunostaining for SCGB3A2 (left panels) and SSEA-1 (right panels) (red) in combination with YFP (green) in control (top panels) and experimental (bottom panels) mice. **h**, Immunostaining for AcTub (left panels) and c-MYB (right panels) (red) in combination with YFP (green) in control (top panels) and experimental (bottom panels) mice. **i**, Flow cytometry analysis of EPCAM$^+$CD24$^+$ ciliated cells and EpCAM$^+$SSEA-1$^+$ secretory cells in control and experimental mice. **j**, Percentage of epithelial (EpCAM$^+$) basal, secretory and ciliated cells from both groups assessed by flow cytometry ($n = 3$ mice). **k**, Immunostaining for p63 (red) on control (top panel) and experimental mice (bottom panel). **l**, Percentage of p63$^+$ cells per total DAPI$^+$ cells on both groups. **m**, Immunostaining for FOXJ1 (green), N2ICD (red) and SCGB1A1 (cyan). **n**, Immunostaining for BrdU (green), p63 (red) and Ki67 (cyan) in either control (upper panels) or experimental mice (lower panels). **o**, Percentage of ciliated FOXJ1$^+$ cells that incorporate BrdU after continuous administration of BrdU to Tam-treated CK5-Jag2$^{fl/fl}$ mice ($n = 3$ mice). **p**, Immunostaining for FOXJ1 (green) and BrdU (red) on Tam-treated CK5-Jag2$^{fl/fl}$ mice that received continuous BrdU ($n = 3$ mice). **q**, Immunostaining to detect apoptotic cells by TUNEL assay (green) on both groups. **r**, Immunostaining for YFP (green) in combination with activated caspase3 (red) on control (upper panel) or experimental mice (lower panel). **f**–**r**, Analysis performed 10 days after induction. Images are representative of $n = 5$ mice per condition (biological replicates) repeated three times. $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$. Data shown in the graphs are means ± s.e.m. Nuclei, DAPI (blue). Scale bar, 20 μm.

doi:10.1038/nature14673

Expression of barley SUSIBA2 transcription factor yields high-starch low-methane rice

J. Su[1,2]*, C. Hu[1,2]*, X. Yan[2]*, Y. Jin[2,3], Z. Chen[1], Q. Guan[1], Y. Wang[1], D. Zhong[1], C. Jansson[4], F. Wang[1], A. Schnürer[5] & C. Sun[2]

Atmospheric methane is the second most important greenhouse gas after carbon dioxide, and is responsible for about 20% of the global warming effect since pre-industrial times[1,2]. Rice paddies are the largest anthropogenic methane source and produce 7–17% of atmospheric methane[2,3]. Warm waterlogged soil and exuded nutrients from rice roots provide ideal conditions for methanogenesis in paddies with annual methane emissions of 25–100-million tonnes[3,4]. This scenario will be exacerbated by an expansion in rice cultivation needed to meet the escalating demand for food in the coming decades[4]. There is an urgent need to establish sustainable technologies for increasing rice production while reducing methane fluxes from rice paddies. However, ongoing efforts for methane mitigation in rice paddies are mainly based on farming practices and measures that are difficult to implement[5]. Despite proposed strategies to increase rice productivity and reduce methane emissions[4,6], no high-starch low-methane-emission rice has been developed. Here we show that the addition of a single transcription factor gene, barley *SUSIBA2* (refs 7, 8), conferred a shift of carbon flux to *SUSIBA2* rice, favouring the allocation of photosynthates to aboveground biomass over allocation to roots. The altered allocation resulted in an increased biomass and starch content in the seeds and stems, and suppressed methanogenesis, possibly through a reduction in root exudates. Three-year field trials in China demonstrated that the cultivation of *SUSIBA2* rice was associated with a significant reduction in methane emissions and a decrease in rhizospheric methanogen levels. *SUSIBA2* rice offers a sustainable means of providing increased starch content for food production while reducing greenhouse gas emissions from rice cultivation. Approaches to increase rice productivity and reduce methane emissions as seen in *SUSIBA2* rice may be particularly beneficial in a future climate with rising temperatures resulting in increased methane emissions from paddies[9,10].

High-starch content and low-methane emissions are two important traits in rice breeding that are difficult to achieve simultaneously. In 2002, high-yielding rice cultivars with improved productivity were proposed as a strategy for reducing methane emissions[4,6]. However, no such high-starch low-methane-emission rice has as yet been reported. Here we report the first example, to our knowledge, of such a rice, *SUSIBA2* rice, generated via transcription factor technology.

Sugar signalling in barley 2 (SUSIBA2) is a plant-specific transcription factor[7,8,11] that regulates sugar-inducible gene expression, thereby mediating source–sink communication[7,8]. High expression of *SUSIBA2* is associated with an increase in sink strength and starch biosynthesis[7,8]. We hypothesized that the overexpression of *SUSIBA2* in the seeds and stems of rice would increase sink strength in aboveground tissues and generate a high-starch low-methane-emission rice variety.

Two stable rice lines (numbers 77 and 80) of homozygote transformants were selected in this study and defined as *SUSIBA2* rice (*SUSIBA2-77* and *SUSIBA2-80*, respectively). *SUSIBA2-77* and its control, Nipponbare (Nipp), were cultivated in Fuzhou, China during the summers of 2012 and 2013. Results showed that cultivation of *SUSIBA2-77* cut methane emissions to around 10% of control levels before flowering, and almost to zero (0.3% of the control level) at 28 days after flowering (Fig. 1a). Genomic sequencing demonstrated that the observed emission trait was related to the introduced *HvSUSIBA2* activity rather than its physical insertion site in the rice genome (Extended Data Fig. 1). Subsequent phytotron experiments corroborated the significant reduction of methane emissions in *SUSIBA2* rice and showed that the measured methane emissions occurred in a linear fashion over time (Extended Data Fig. 2).

To illustrate the role of *SUSIBA2* rice in reducing methane emissions in different ecological environments and climates, we cultivated *SUSIBA2-77* and *SUSIBA2-80* in the autumn of 2014 in Fuzhou, Guangzhou and Nanning, China at locations >500 km apart, and measured diurnal and seasonal methane emissions (Extended Data Fig. 3). Both rice lines displayed similar traits in methane emissions, that is, less reduction in the morning, but significant reduction during the day (Extended Data Fig. 3). Interestingly, the observed pattern of emission reduction matched the predicted pattern of SUSIBA2-controlled sugar metabolism, which increases during the day and in the summer.

To elucidate the mechanism underpinning the reduced methane emissions from *SUSIBA2* rice, we took a three-pronged approach: (1) cultivating *SUSIBA2* and control rice in phytotrons/fields and quantifying rhizospheric methanogen communities; (2) measuring phenotypic traits of *SUSIBA2* rice; and (3) characterizing genotypic traits.

Strong blue-green autofluorescence when excited with light at 420 nm wavelength is a characteristic of all methanogens due to the presence of the cofactor F420 (ref. 12). Fluorescence microscopy revealed that the major methanogenic autofluorescence was associated with the root tip and proximal regions, and much less signal was apparent in *SUSIBA2-77* compared with Nipp (Fig. 1b). Quantification of the methanogenic communities indicated that the gene copy numbers of total archaea and methanogens and the orders Methanobacteriales, Methanomicrobiales and Methanocellales and two families Methanosaetaceae and Methanosarcinaceae of the order Methanosarcinales were significantly lower in *SUSIBA2* rice compared with Nipp in both phytotron (Fig. 1c and Extended Data Fig. 4) and field conditions (Extended Data Fig. 5). We propose that the low-methane emissions from *SUSIBA2* rice were due to a decrease in the rhizospheric abundance of methanogens. Methanocellales are the dominant methanogens in rice paddies[13,14], although recent reports indicate that other methanogenic groups can also be abundant in certain paddies[15]. One plausible reason for the broad representation of methanogens detected in this study could be that plant residues were included as fertilisers in field and phytotron soils, possibly favouring richness in the methanogenic population[16,17]. When we looked into the

[1]Institute of Biotechnology, Fujian Academy of Agricultural Sciences, Fuzhou 350003, China. [2]Department of Plant Biology, Uppsala BioCenter, Linnean Center for Plant Biology, Swedish University of Agricultural Sciences, PO Box 7080, SE-75007 Uppsala, Sweden. [3]Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization, Hunan Agricultural University, Changsha 410128, China. [4]The Environmental Molecular Sciences Laboratory (EMSL), Pacific Northwest National Laboratory, PO Box 999, K8-93 Richland, Washington 99352, USA. [5]Department of Microbiology, Uppsala BioCenter, Swedish University of Agricultural Sciences, SE-75007 Uppsala, Sweden.
*These authors contributed equally to this work.

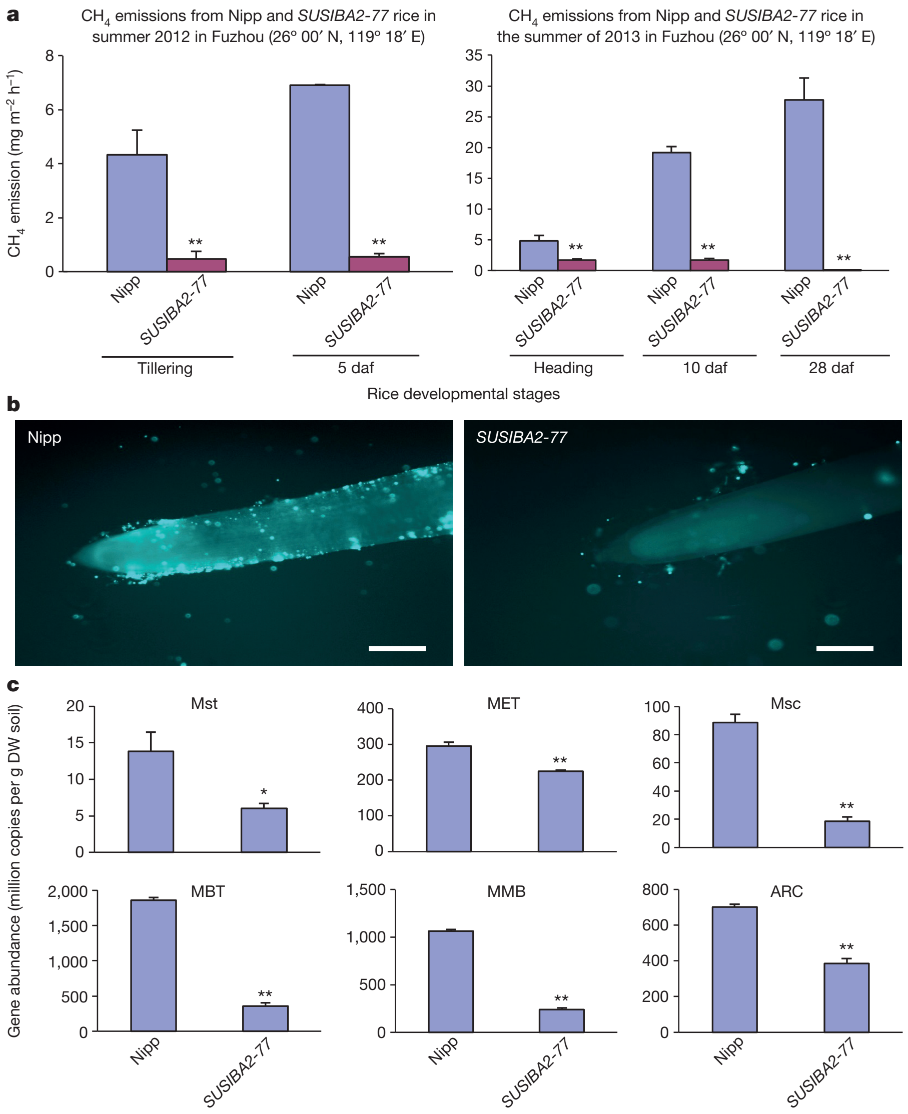


Figure 1 | *SUSIBA2* rice reduces methane emissions from rice paddies. **a**, Methane emissions from four rice plants ($n = 4$) for 2013 and three ($n = 3$) for 2012. **b**, Fluorescent microscopy of Nipp and *SUSIBA2-77* roots. Scale bar, 1 mm. **c**, Quantification of methanogens in soil and root samples from three rhizospheric positions ($n = 3$) of three plants ($n = 3$). Technical triplicates per position were applied. Quantification was performed for total archaea (ARC) and methanogens (MET), and the orders Methanobacteriales (MBT), Methanomicrobiales (MMB) and Methanocellales and two families, Methanosaetaceae (Mst) and Methanosarcinaceae (Msc), of the order Methanosarcinales, respectively. Typical results from soil sample 1 of plant 1 are shown. Differences between Nipp and *SUSIBA2-77* were statistically significant (one-way ANOVA, $*P \leq 0.05$ or $**P \leq 0.01$, error bars show s.d.). daf, days after flowering; DW, dry weight.

specific group of Methanocella[14] in the order of Methanocellales, we found that it was also significantly reduced in *SUSIBA2* rice (Extended Data Fig. 5c).

To investigate if photosynthate partitioning in *SUSIBA2* rice was altered, we examined phenotypic traits in *SUSIBA2-77* and Nipp. Compared with Nipp, *SUSIBA2-77* has larger panicles with a higher proportion of filled grains, resulting in more drooping panicles (Fig. 2a, left and middle panels). In contrast, *SUSIBA2-77* has a smaller root system than Nipp (Fig. 2a, right panel). Consistent with these observations, measurements of total dry biomass in aboveground and belowground tissues (Fig. 2b) showed that significantly more aboveground biomass and less root biomass were found with *SUSIBA2-77*. The increased aboveground biomass was associated with grain dry biomass and the numbers of filled grains (Fig. 2b). Plant height, thousand-grain weight (TGW), and numbers of panicles and tillers per plant were similar in *SUSIBA2-77* and Nipp (Fig. 2b).

Starch content in filled grains of *SUSIBA2-77* increased to 86.9% dry weight compared with 76.7% in Nipp (Fig. 2c). Electron microscopy examination showed that the starch granule sizes in mature seeds of *SUSIBA2-77* were reduced (Fig. 2d). As the grain size was unaffected in *SUSIBA2-77*, we suggest that one possible reason for the increased starch content in mature *SUSIBA2-77* grains is a more dense packing of smaller granules. A closer examination of starch content during seed development showed an elevation from 14 days after flowering (daf) (Fig. 2c). Notably, at the same developmental stage (14 daf), the starch content in stems was also significantly increased, but not in leaves and roots (Fig. 2c).

To link the phenotypic traits with genotypic profiles, we examined the integration of *HvSUSIBA2* in the *SUSIBA2* rice genome, along with the gene and protein expression of *HvSUSIBA2* and other representative genes. Southern blot analysis revealed two copies of *HvSBEIIb* p:*HvSUSIBA2* in the *SUSIBA2-77* and *SUSIBA2-80* genome (Fig. 3a), in agreement with the sequencing results from *SUSIBA2-77* (Extended Data Fig. 1b). The presence of sugar-responsive elements (SUREs) in the promoters of genes targeted by SUSIBA2 is critical for executing the sugar-signalling cascade that controls starch biosynthesis[7,8]. Using

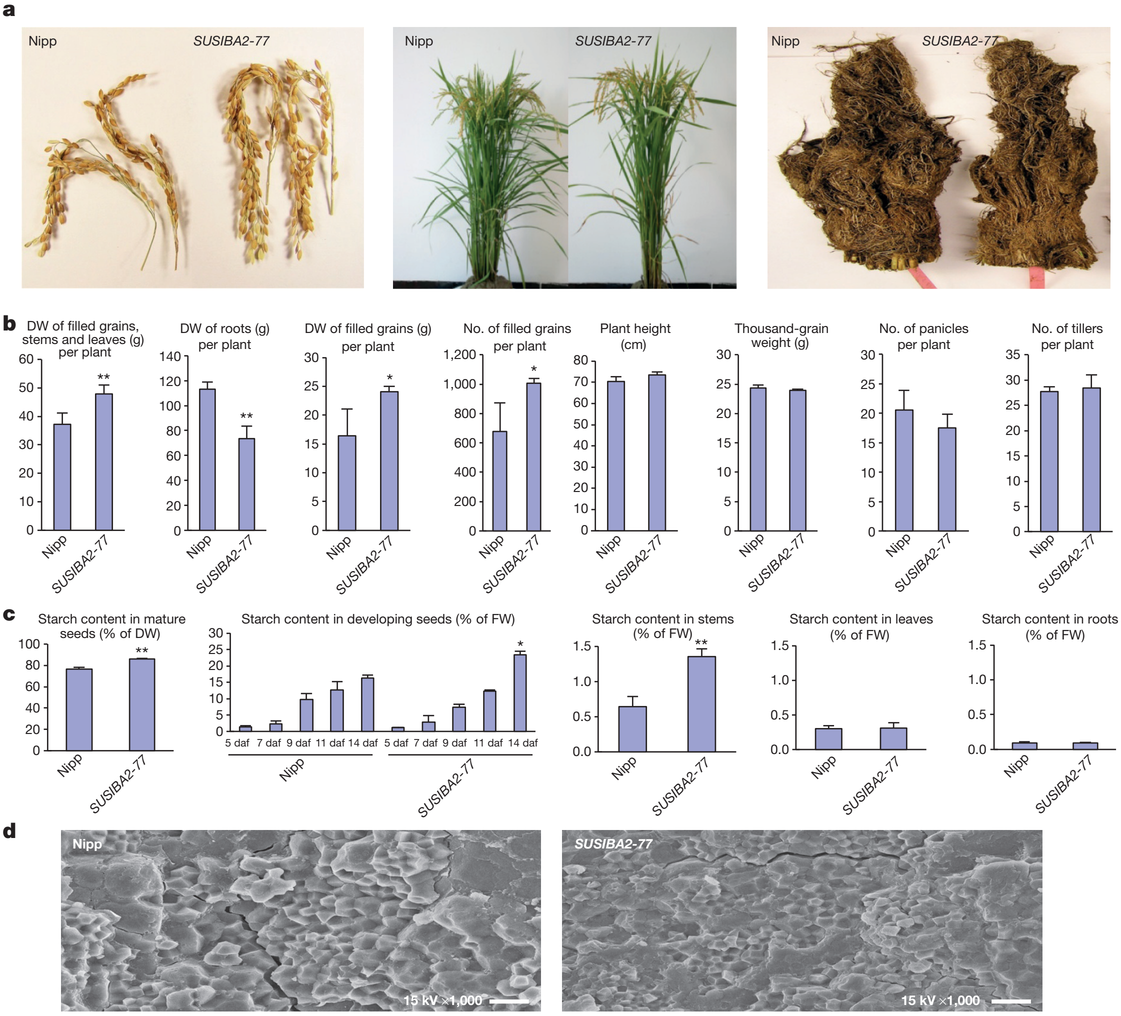


Figure 2 | Phenotypic profiling of *SUSIBA2* rice and Nipponbare (Nipp). **a**, Panicles (left), aboveground plants (middle) and roots (right). **b**, For *SUSIBA2-77* we observed significantly increased dry weight of aboveground biomass and filled grains, increased number of filled grains, and significantly decreased root dry weight. Plant height, thousand-grain weight, number of panicles and tillers were not significantly altered in *SUSIBA2-77*. Four plants were used ($n = 4$). **c**, For *SUSIBA2-77* starch content in mature seeds, developing seeds and stems at 14 daf changed significantly, but did not change in leaves and roots. Three plants ($n = 3$) were used. **d**, Starch granules from mature seeds. Scale bar, 10 μm. One-way ANOVA was used for statistical analysis (*$P \leq 0.05$ or **$P \leq 0.01$, error bars show s.d.). DW, dry weight; FW, fresh weight.

the protocol described in ref. 7, we demonstrated the SURE-binding activity in rice by HvSUSIBA2 (Extended Data Fig. 6). To investigate gene expression, we selected 24 genes[18] associated with sugar metabolism, including *HvSUSIBA2* and *OsSUSIBA2-like*. Gene expression analyses showed that *HvSUSIBA2* was highly expressed in early developing seeds and stems, and at very low levels in leaves, roots and late developing seeds (Fig. 3b, upper two panels). The expression pattern correlated with the activity of the *HvSBEIIb* promoter in rice (Extended Data Fig. 7). The expression level of selected genes, apart from the control gene *TIP41-like*, followed the *HvSUSIBA2* expression pattern with significant differences between *SUSIBA2* rice and Nipp in seeds (Fig. 3b and Extended Data Figs 8 and 9) and, at least for some genes, in stems but not in leaves or roots. (Fig. 3b, Extended Data Fig. 8 and Supplementary Table 2). Following transcriptomic analysis, the same tissues of *SUSIBA2-77* were subjected to protein expression analysis. Protein analysis indicated that all five selected proteins could be found at higher levels in stems, but not in leaves and roots in *SUSIBA2-77* compared with Nipp (Fig. 3c, upper panel). In developing seeds, the membrane protein (SUT5) and starch granule-bound protein (GBSSI) were expressed more in *SUSIBA2-77* than in Nipp (Fig. 3c, lower panel). The same trend was not obvious for the soluble proteins. *In vitro* degradation experiments suggested that the unchanged levels of soluble proteins in *SUSIBA2-77* compared with Nipp might represent a higher turnover of soluble proteins in *SUSIBA2-77* (Fig. 3d), supported by a zymogram activity assay of a soluble phase of starch branching enzyme I (Fig. 3e).

We have proposed a model for how *SUSIBA2* rice works (Fig. 4). The *HvSBEIIb* promoter activity is sugar-inducible via the activation of *HvSUSIBA2* (refs 7, 8). When photosynthates (sugars) are available, *HvSBEIIb* p:*HvSUSIBA2* enhances sugar-inducible activities of targeted genes, including the *HvSBEIIb* p construct. The augmented gene expression increases sink strength in tissues where *HvSUSIBA2* is expressed. The increased sink strength draws more sugars from source tissues to yield more biomass and more filled grains associated with higher starch content. The increased allocation of photosynthates further activates *HvSBEIIb* p:*HvSUSIBA2*, causing a ‘snowball effect’ for increasing sink strength. The snowball effect was illustrated by the exogenous sugar treatment of leaves and field trial experiments. Expression levels of the sugar-inducible genes were significantly higher in *SUSIBA2* rice than in Nipp when sucrose was available (Extended Data Fig. 10). Methane emission reduction was more effective during the summer than during the autumn, and at noon rather than in the morning or later afternoon (Fig. 1a and Extended Data Fig. 3). The data indicate that the higher temperature and enhanced sugar metabolism during the summer and at noon may favour the SUSIBA2-derived snowball effect and carbon allocation to aboveground biomass. As suggested in the model, less sugar

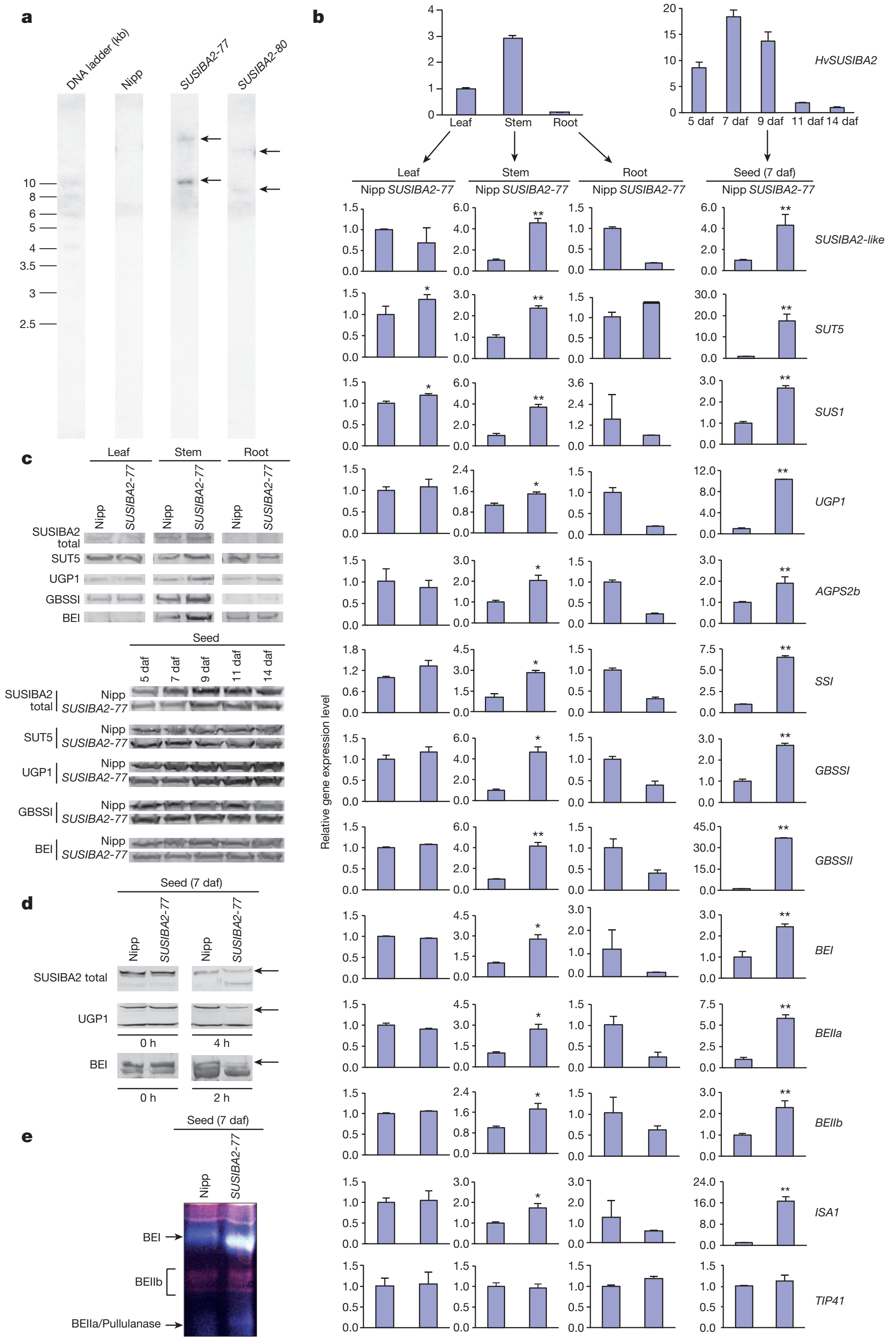


Figure 3 | Genotypic profiling of *SUSIBA2* rice and Nipponbare (Nipp). **a**, Two copies of *HvSUSIBA2* were detected in *SUSIBA2-77* and *SUSIBA2-80*. **b**, Downstream effects of *HvSUSIBA2* expression in *SUSIBA2-77*. High expression of *HvSUSIBA2* enhanced expression of 12 genes in stems and early developing seeds, with a corresponding reduction of gene expression in roots. Three plants ($n = 3$) and technical triplicates per plant were used. **c**, Western blot analysis. SUSIBA2 total represents levels of HvSUSIBA2 and OsSUSIBA2-like. Protein levels increased in stems of *SUSIBA2-77*, but not in leaves and roots (upper panel). Protein levels also increased in seeds for membrane proteins (SUT5) and starch granule-bound proteins (GBSSI), but not for soluble proteins (lower panel). **d**, *In vitro* study of degradation efficiency of soluble proteins in *SUSIBA2-77* seeds. A higher degradation efficiency was detected in *SUSIBA2* rice. **e**, Zymogram of starch branching enzyme I (BEI) activity. Soluble branching enzyme I (BEI) activity increased in *SUSIBA2-77*. One-way ANOVA was used for statistical analysis (*$P \leq 0.05$ or **$P \leq 0.01$, error bars show s.d.).

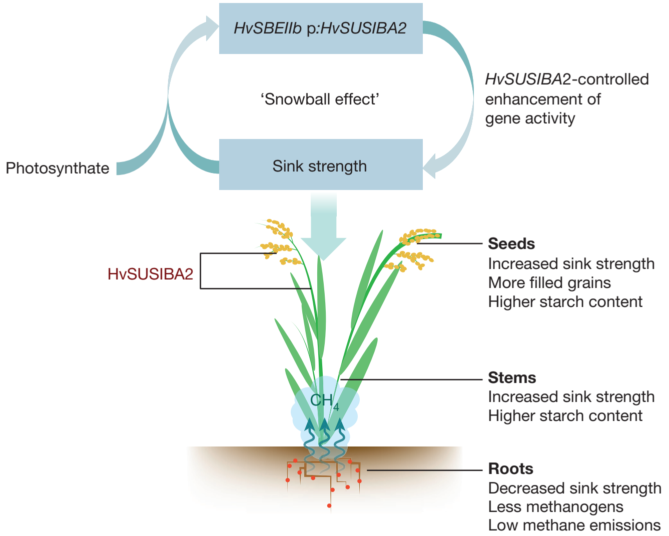


Figure 4 | Model depicting high-starch low-methane-emission *SUSIBA2* rice. Sugar-inducible *HvSUSIBA2* expression generates a snowball effect that ultimately leads to a rice plant with enhanced starch accumulation in seeds and stems and decreased carbon allocation to roots, which reduces methanogenic growth and methane emissions. Red dots represent methanogens.

availability in *SUSIBA2* rice roots results in a rhizosphere with less organic biomass and hence less root exudates as nutrients for inhabiting methanogenic consortia. As a consequence, methane emissions decreased. Due to the biomass differences between *SUSIBA2* rice and Nipp, we do not exclude the possibility that the emission reduction is due in part to impaired methane transport in *SUSIBA2* rice.

Finally, we suggest that the use of *SUSIBA2* rice in cutting methane emissions from paddies may become more relevant with global warming. Increased temperatures accelerate methane emissions in all ecosystems including paddies[9,10], but they also favour SUSIBA2-derived carbon allocation to seeds and aboveground biomass, thus counteracting the temperature-driven acceleration of rice paddy methane emissions.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 June 2014; accepted 19 June 2015.
Published online 22 July 2015.

1. Kirschke, S. *et al.* Three decades of global methane sources and sinks. *Nature Geosci.* **6,** 813–823 (2013).
2. Bridgham, S. D., Hinsby, C.-Q., Jason, K. K. & Zhuang, Q. Methane emissions from wetlands: biogeochemical, microbial, and modeling perspectives from local to global scales. *Glob. Change Biol.* **19,** 1325–1346 (2013).
3. Liu, Y. & Whitman, W. B. Metabolic, phylogenetic, and ecological diversity of the methanogenic archaea. *Ann. NY Acad. Sci.* **1125,** 171–189 (2008).
4. Sass, R. L. & Cicerone, R. J. Photosynthate allocations in rice plants: Food or atmospheric methane. *Proc. Natl Acad. Sci. USA* **99,** 11993–11995 (2002).
5. Qiu, J. China cuts methane emissions from rice fields. *Nature.* http://www.nature.com/news/2009/090818/full/news.2009.833.html (2009).
6. Denier van der Gon, H. A. *et al.* Optimizing grain yields reduces CH_4 emission from rice paddy fields. *Proc. Natl Acad. Sci. USA* **99,** 12021–12024 (2002).
7. Sun, C. *et al.* A novel WRKY transcription factor, SUSIBA2, participates in sugar signaling in barley by binding to the sugar-responsive elements of the *iso1* promoter. *Plant Cell* **15,** 2076–2092 (2003).
8. Sun, C., Höglund, A.-S., Olsson, H., Mangelsen, E. & Jansson, C. Antisense oligodeoxynucleotide inhibition as a potent strategy in plant biology: identification of SUSIBA2 as a transcriptional activator in plant sugar signaling. *Plant J.* **44,** 128–138 (2005).
9. Yvon-Durocher, G. *et al.* Methane fluxes show consistent temperature dependence across microbial to ecosystem scales. *Nature* **507,** 488–491 (2014).
10. Hoehler, T. M. & Alperin, M. J. Methane minimalism. *Nature* **507,** 436–437 (2014).
11. Rushton, P. J., Somssich, I. E., Ringler, P. & Shen, Q. J. WRKY transcription factors. *Trends Plant Sci.* **15,** 247–258 (2010).
12. Ashby, K. D., Casey, T. A., Rasmussen, M. A. & Petrich, J. W. Steady-state and time-resolved spectroscopy of F420 extracted from methanogen cells and its utility as a marker for fecal contamination. *J. Agric. Food Chem.* **49,** 1123–1127 (2001).
13. Liu, P., Yang, Y., Lü, Z. & Lu, Y. Response of a rice paddy soil methanogen to syntrophic growth as revealed by transcriptional analyses. *Appl. Environ. Microbiol.* **80,** 4668–4676 (2014).
14. Angel, R., Claus, P. & Conrad, R. Methanogenic archaea are globally ubiquitous in aerated soils and become active under wet anoxic conditions. *ISME J.* **6,** 847–862 (2012).
15. Edwards, J. *et al.* Structure, variation, and assembly of the root-associated microbiomes of rice. *Proc. Natl Acad. Sci. USA* **112,** E911–E920 (2015).
16. Conrad, R., Klose, M., Lu, Y. & Chidthaisong, A. Methanogenic pathway and archaeal communities in three different anoxic soils amended with rice straw and maize straw. *Frontiers Microbiol.* **3,** http://dx.doi.org/10.3389/fmicb.2012.00004 (2012).
17. Peng, J., Lü, Z., Rui, J. & Lu, Y. Dynamics of the methanogenic archaeal community during plant residue decomposition in an anoxic rice field soil. *Appl. Environ. Microbiol.* **74,** 2894–2901 (2008).
18. Zhang, M.-Z. *et al.* Molecular insights into how a deficiency of amylose affects carbon allocation-carbohydrate and oil analysis and gene expression profiling in the seeds of a rice waxy mutant. *BMC Plant Biol.* **12,** 230 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements Special thanks to S. Stymne. We would also like to thank B. Müller, X. Feng, M. Erikson, L. Sun, S. Isaksson, J. Ascue and S. Mayer for their help in determining concentrations of methane and methanogens, and B. Ingemarsson for discussions concerning the work layout. This work was funded by the following organisations and foundations: The Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (Formas) for Project No 219-2014-1172; the joint Formas/Sida-funded programme (Project No 220-2009-2069) on sustainable development in developing countries; the SLU Lärosätesansökan Programme (TC4F) for Team 4 supported by Vinnova; the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (Formas) under the Strategic Research Area for the TCBB Programme; National Natural Science Foundation of China (projects no 30771298 and no 31370389); the SLU programme BarleyFunFood; the Carl Trygger Foundation for Project No CTS 11: 450; funding in part by the US Department of Energy Contract DE-AC05-76RL01830 with the Pacific Northwest National Laboratory.

Author Contributions J.S., Z.C., Q.G., Y.W. and D.Z. performed measurements of methane emissions from paddies; J.S. also performed western blot and zymogram analyses, methanogen quantification and starch determination. C.H. was responsible for plasmid constructions, rice transformation, Southern blot analysis and phenotypic trait characterization. X.Y. carried out gene expression analysis, starch determination, sugar induction experiments and phenotypic trait characterization. Y.J. performed plasmid validation, insertion site identification, methanogen quantification, measurements of methane emissions in phytotrons and sugar induction experiments, electrophoretic mobility shift assay (EMSA), qPCR and light microscopy. C.J. was involved in the initiation, layout and discussions concerning the work and manuscript revision. F.W. was involved in the planning of rice transformation and field trial settings. A.S. revised the manuscript and helped with methane and methanogen determinations. C.S. initiated and coordinated the work, designed the experiments, performed some experiments, and drafted and revised the manuscript.

Author Information The sequence of construct containing *HvSBEIIb* p:*HvSUSIBA2* has been deposited in GenBank under accession number KR935231. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.S. (Chuanxin.Sun@slu.se) or F.W. (wf@fjage.org).

METHODS

Plant materials and growth conditions. Rice plants of variety Nipponbare (*Oryza sativa* L. ssp. *Japonica*) and transformed homozygote lines *SUSIBA2-77* and *SUSIBA2-80* were grown in open fields or in a phytotron. Open field cultivation was performed in a similar way to that described previously[18], but with rice straw in soil under natural conditions in Fuzhou, Guangzhou and Nanning, southern China, respectively. Phytotron conditions were applied to mimic field conditions, but with limited high temperatures. In the phytotron, rice plants were grown in cylinder-type pots (30 cm high with an upper diameter of 29 cm and bottom diameter of 19 cm) with organic soil containing plant residues. Phytotron growth management was similar to that described previously[19] with a modified setting for rice, 14 h light/10 h dark at 30 °C/21 °C, a constant relative humidity of 80% and light intensity of 400 µmol photons m^{-2} s^{-1}.

Plasmid construction and rice transformation. Plasmid construction and general molecular cloning procedures were performed according to previously developed protocols[7,8,20]. Nucleotides 247–2067 of GenBank accession number AY323206, encoding barley *SUSIBA2*, were fused to nucleotides 1–1010 of barley *SBEIIb* promoter (*HvSBEIIb* p; GenBank accession number AF064563). The fused DNA fragment was cloned in the pCAMBIA 1301 binary vector (Extended Data Fig. 1a) and sequenced at Macrogen Europe (Amsterdam, the Netherlands). The sequence of the construct is in the Supplementary Information and deposited under GenBank accession number KR935231. The plasmid construct was used for *Agrobacterium*-mediated transformation of rice following the protocol in ref. 21. Screening of post-transformants was based on hygromycin resistance and PCR determination of T-DNA insertion. Out of 14 positive lines, five homozygous lines were selected for characterization. Eventually, two homozygote lines *SUSIBA2-77* (Extended Data Fig. 1b) and *SUSIBA2-80* were used for detailed studies. A binary vector containing *HvSBEIIb* p:*GUS* was also constructed and transformed to Nipponbare (Extended Data Fig. 7). The final construct was verified by DNA sequencing, and transformed into *Agrobacterium tumefaciens* strain EHA105.

Gene expression analysis by quantitative PCR (qPCR). RNA isolation, cDNA synthesis and qPCR analysis were performed in accordance with previous reports[8,18]. In brief, plant materials from different tissues were ground into fine powders in liquid nitrogen and total RNA was isolated by the Spectrum Plant Total RNA Kit (Sigma-Aldrich) according to the manufacturer's protocol using approximately 30 mg of plant materials. All samples were treated with DNase I (Sigma-Aldrich) to remove trace amounts of DNA contamination. Total RNA of 1 µg was used as a template for the cDNA synthesis with the Quanta qScript cDNA synthesis kit (Quanta Biosciences). The synthesized cDNA was adjusted to a concentration of 5 ng µl^{-1} and 15 ng used for qPCR analysis. qPCR reactions with at least 90% amplification efficiency were performed in a volume of 20 µl containing 5 µM specific primers and a SYBR Green PCR master mix (Applied Biosystems, Life Technologies Europe BV). The PCR programme consisted of an initial temperature of 95 °C for 4 min, and then 40 cycles of 30 s at 95 °C and 30 s at 60 °C. The melt curve was performed by increasing the temperature from 60 °C to 95 °C at a speed of 0.05 °C per second. qPCR-specific amplification was verified by a single band product in gel analysis. Data were calculated with the comparative C_t method[18] and one-way ANOVA[18] was used for statistical analysis. The gene expression level by qPCR was normalized using housekeeping genes *ACT11* (ref. 22) in developing seeds and *Ubiquitin10* (ref. 22) in vegetative tissues, and using *TIP41-like*[23] as a control gene.

Electrophoretic mobility shift assay (EMSA). Overexpressed HvSUSIBA2 protein from *E. coli* was purified and used for EMSA. EMSA was performed essentially as described in ref. 7. The SURE sequences in the rice *ISA1* promoter were found by manual search in the promoter sequence[18]. The binding assay of HvSUSIBA2 protein to labelled SURE oligonucleotides was performed by incubating the protein with a DNA probe at room temperature for 30 min and then visualizing a protein–DNA complex on a native 5% polyacrylamide gel. The gel electrophoresis was conducted at 200 V for 2 h. A DNA fragment of 'A-rich stretch' was used as a negative control. The gel was dried and autoradiography was carried out on an X-ray film overnight.

Exogenous sucrose induction. Leaf blades were excised from rice plants after 10 h in the dark during the early tillering stage and incubated with 100 mM sucrose solutions in a falcon tube (1 blade per tube) containing 2 ml sucrose solution for 24 h at 22 °C in the dark. Immediately after collection, the leaf blades were frozen in liquid nitrogen and stored at −70 °C until further analysis.

Validation of T-DNA insertion sites. The genomic DNA of *SUSIBA2-77* was isolated using a DNA isolation kit (Qiagen GmbH, Hilden, Germany). The genomic DNA was fragmented by HindIII and ligated to a DNA adaptor also generated by HindIII. The HindIII-digested adaptor was purified before ligation using a PCR product purification kit (Qiagen). PCR cloning was performed using the ligated template to obtain a specific PCR product with the primers against T-DNA board sequences and the adaptor sequences. The resulting PCR products were then sub-cloned into a PCR cloning vector by the TOPO10 TA cloning kit (Life Technologies Europe BV) for sequencing.

qPCR quantification of methanogenic communities. Soil or root samples from three independent positions in the underground vicinity around the root tip and proximal regions (5 cm from the rice plant and 5 cm depths, and 5 cm between positions) of three independent plants for Nipp, *SUSIBA2-77* and *SUSIBA2-80* were collected at 3.00 p.m. from rice paddies or phytotrons, and the DNA was subsequently isolated using a DNA isolation kit for soil organisms (FastDNA SPIN Kit for Soil; MP Biomedicals, LLC). The DNA was then diluted into four different series concentrations. qPCR quantification was performed for all four diluted DNA samples using a standard of a previously cloned 16S rRNA gene fragment[24–26] for the individual groups of targeted methanogens and a newly PCR cloned 16S rRNA gene fragment using described primers[14] from the strain *Methanocella conradii* DSM 24694 (German Collection of Microorganisms and Cell Cultures GmbH, Germany) for the genera of *Methanocella*. Only DNA copy numbers that were in agreement from two series of diluted samples were used for copy number determination. All methanogenic groups were analysed by using group-specific primers (Supplementary Table 1) with at least 90% amplification efficiency. Existing PCR products in all qPCR reactions were verified by gel analysis, showing a single band of the expected size. The abundance of each group was calculated and translated to DNA copy numbers for each gram of dry root and/or soil. The abbreviations for each group are Mst (Methanosaetaceae), Msc (Methanosarcinaceae), MBT (Methanobacteriales), MMB (Methanomicrobiales), ARC (archaea) and MET (methanogens) according to ref. 26, and Met for *Methanocella-specific* (Supplementary Table 1). The qPCR programme for Mst was as follows: 95 °C for 7 min, then 54 cycles of 40 s at 95 °C, 1 min at 61 °C and 40 s at 72 °C; for MBT: 95 °C for 7 min, followed by 54 cycles of 40 s at 95 °C, 1 min at 58 °C and 40 s at 72 °C; for MMB: 95 °C for 7 min, followed by 54 cycles of 40 s at 95 °C, 1 min at 66 °C and 40 s at 72 °C; for Msc, ARC and MET: 95 °C for 7 min, followed by 54 cycles of 40 s at 95 °C, 1 min at 60 °C and 40 s at 72 °C; for *Methanocella*-specific (Met) analysis: 94 °C for 4 min, followed by 40 cycles of 30 s at 94 °C, 1 min at 60 °C. All melting curves are from 55 °C to 95 °C with an increase of 0.05 °C per second.

Autofluorescence analysis of methanogens. Rice root tip and proximal regions at/after 28 days after flowering were picked at 3.00 p.m. and rinsed with pure water and then examined under a microscope at an excitation light wavelength of 420 nm. Multiple observations on different batches of rice plants were performed.

Southern blot analysis. Southern blot analysis was performed according to a previous protocol[20]. Briefly, rice genomic DNA was isolated from leaves using a CTAB method and 10 µg DNA was used for restriction digestion at 37 °C for 2 h with BamHI. Digested DNA was applied to agarose gel separation immediately with an electrophoresis condition of 50 V for about 7 h. The separated DNA was denatured by an alkaline solution and transferred to the Hybond-N membrane overnight through capillary blotting with 10× SSC solution, and then the DNA was cross-linked to the membrane by UV-light for 5 min. The hybridization was performed overnight with a probe labelled with α-^{32}P dCTP using a random prime labelling kit (rediprime II, GE Healthcare) in a volume of 20 ml hybridization solution at 42 °C. After hybridization, the membrane was washed under a moderate stringent condition, 1× SSC at 50 °C. Autoradiography was conducted with an X-ray film on the membrane.

Western blot analysis. Western blot analysis was performed as described previously[7]. Five proteins were selected for analysis: a membrane protein (sucrose transporter 5, SUT5), a starch granule-bound protein (granule-bound starch synthase I, GBSSI), and three soluble/soluble-phase proteins (SUSIBA2 total = barley HvSUSIBA2 and rice OsSUSIBA2-like; UDP-glucose pyrophosphorylase 1, UGP1; branching enzyme I, BEI), respectively. Peptide antibodies against the five proteins were obtained from the Beijing Genomics Institute. The antibodies were raised against peptide positions of 300–330, 510–540, 440–470, 260–290 and 480–510 of GenBank accession numbers NM_001066651, DQ072593, DQ395328.1, X62134 and D11082, for OsSUSIBA2-like, SUT5, UGP1, GBSSI and BEI, respectively. Rice tissue samples were collected at 3.00 p.m. and ground into fine powder. Total proteins were extracted from approximately 200 mg samples using the plant total protein extraction kit (Sigma-Aldrich, St. Louis, MO, US). The proteins were separated in a 4–12% gradient PAGE gel under a condition of 150 V for 1.5–2 h, and then transferred to PVDF membrane by electro-blotting. The PVDF membrane was blocked in 5% milk in a TBS buffer for 1 h, and then incubated with the first antibody in the blocking solution overnight at room temperature. Incubation with the second antibody conjugated with a phosphatase was performed for approximately 2 h after the membrane was washed by TBST and TBS solutions. Immuno-reacted bands were visualized in a BCIP/NBT (substrates for the phosphatase) solution for colour development.

Methane collection in rice paddies and phytotrons. Methane sampling and determination were based on published protocols[27–29]. Individual rice plants were covered with a sealed plastic cylinder (diameter: 15 cm, height: 45 cm for early stages and 95 cm for late stages) and 3 × 25 ml gas samples were taken from six independent plants by a syringe from the headspace after 10, 20 and 30 min and pooled in a sealed airbag. Sampling time in rice paddies was 8.00 a.m. for the summer (June–August) of 2012 and 2013, and 8.00 a.m., 12.00 p.m. and 4.00 p.m. for the autumn (September–November) of 2014, respectively. Temperatures were recorded. The samples in the phytotrons were taken from six independent plants at 3.00 p.m. after 15 min coverage for each plant. Four technical repeats, that is, four vials, were used for each plant methane collection. All gas samples were analysed by gas chromatography with appropriate methane standards. Methane in samples was quantified by calculating the sample peak area in comparison with the standard peak area of known amounts. An air methane concentration of 1.8 p.p.m. was used as the background for calculations. Methane flux calculation was according to the formula described by ref. 29.

Degradation efficiency analysis. Soluble proteins were extracted from the seeds at 7 daf without the addition of protease inhibitors according to a protocol described previously[7,8]. The protein extracts were adjusted to the same concentration at time = 0 and incubated at 28° C for 2 or 4 h. Western blot analysis was used to check the remaining protein levels.

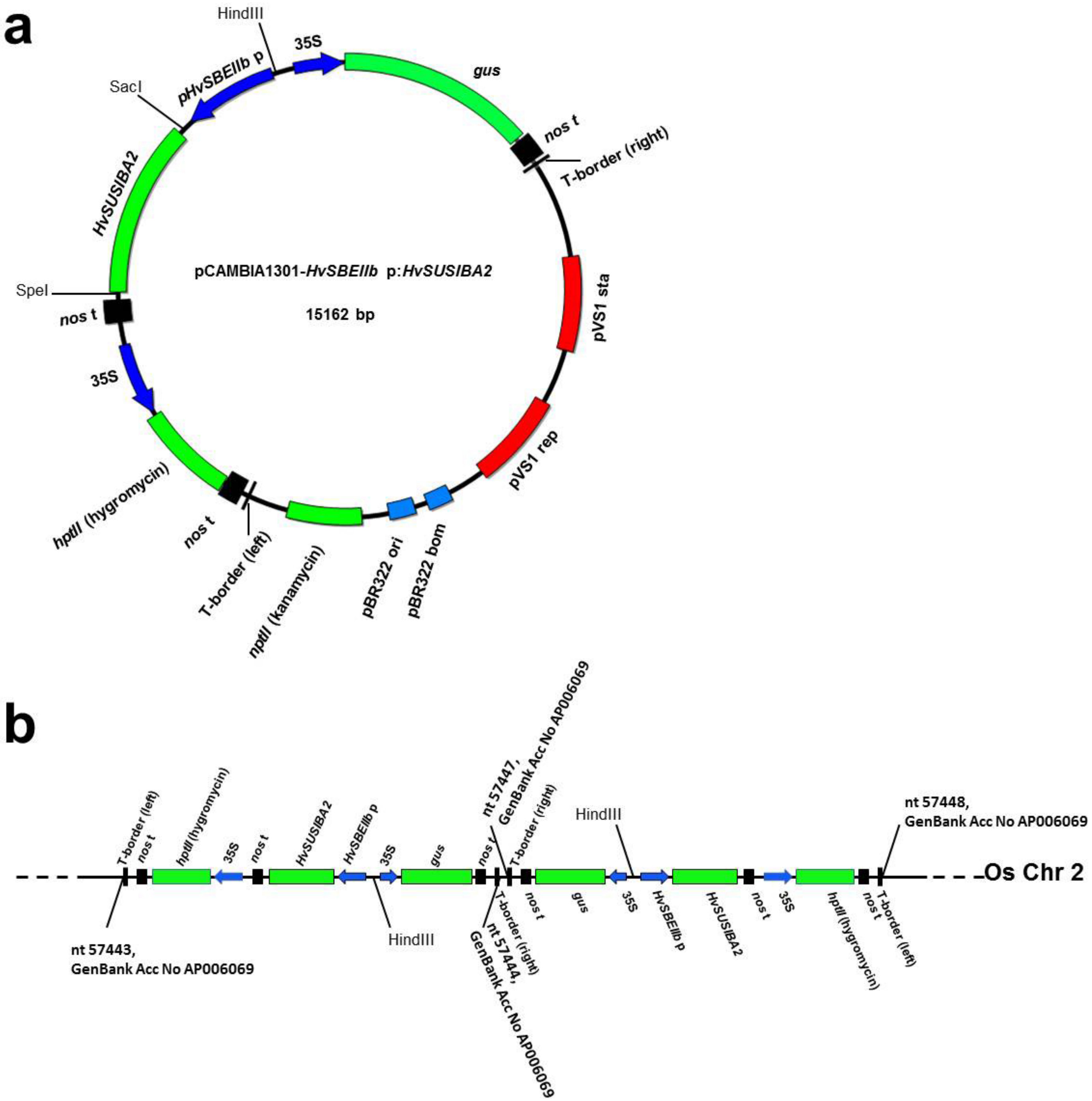
Total starch analysis. Total starch was extracted and analysed using a total starch assay kit (Megazyme, Bray, Co. Wicklow, Ireland). Rice tissues were collected at 3.00 p.m. and ground into fine powder in liquid nitrogen and 100 mg samples of tissues were used for total starch analysis after removing the soluble sugars with 80% ethanol. Total starch quantification and calculation followed the manufacturer's protocol exactly.

Zymogram analysis of starch branching enzyme (SBE) activity. Zymogram analysis was performed according to a protocol described previously[8]. Equal amounts of total soluble proteins from Nipp and *SUSIBA2-77* developing seeds at 7 daf were separated in a 10% (w/v) polyacrylamide gel containing 1% (w/v) starch. After electrophoresis, the gel containing starch branching enzymes was washed with water and incubated at 30 °C in a buffer containing 100 mM sodium citrate (pH 7.0) for 4–8 h. The gel was then transferred to an I_2/Kl solution to visualize reddish-purple bands against the dark blue starch gel for SBE activity.

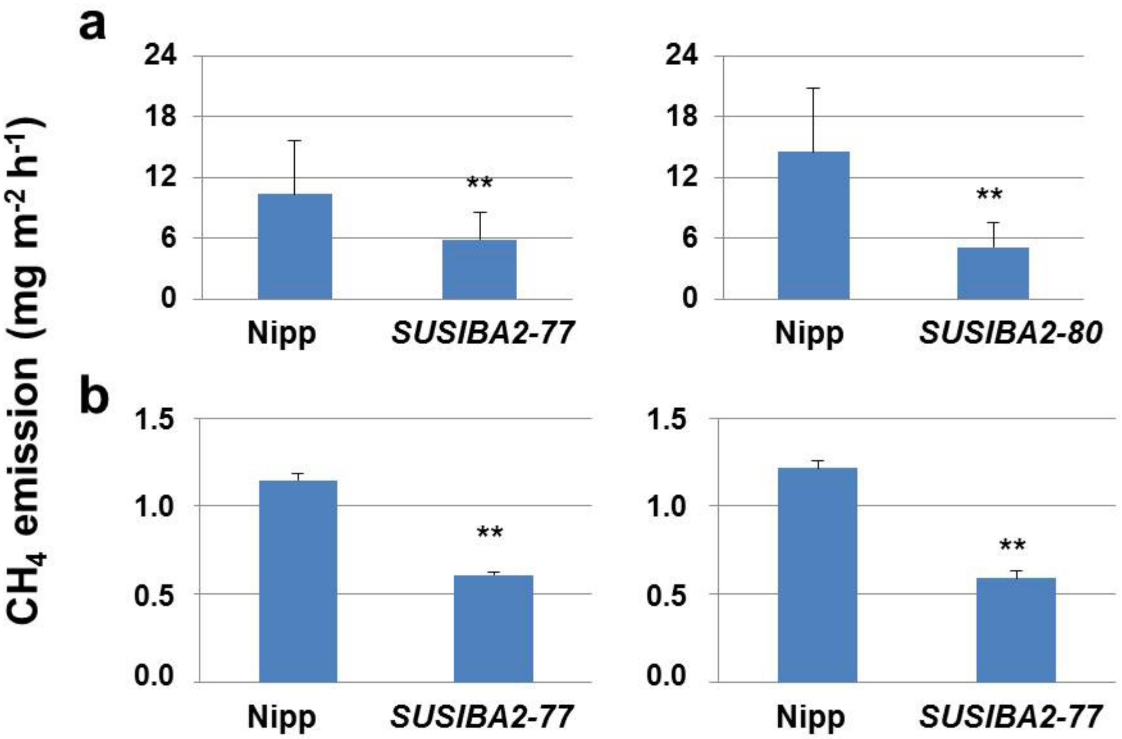
Scanning electron microscopy of starch granules. Starch granules of Nipponbare and *SUSIBA2-77* were isolated as described[8]. Coating of the granules was carried out in a high-resolution sputter coater, with platinum/palladium as the target. Scanning electron microscopy was performed on a JSM-6320F (JEOL Ltd, Akishima, Tokyo, Japan).

Molecular cloning of *Methanocella*-specific 16S rRNA genes. Genomic DNA of *Methanocella conradii* DSM 24694 were isolated by the DNA isolation kit for soil organisms (FastDNA SPIN Kit for Soil; MP Biomedicals, LLC) and used as PCR templates for PCR-based cloning of 16S rRNA. PCR-amplified products were subcloned into a PCR cloning vector by the TOPO10 TA cloning kit (Life Technologies Europe BV). The cloned genes were verified by DNA sequencing.

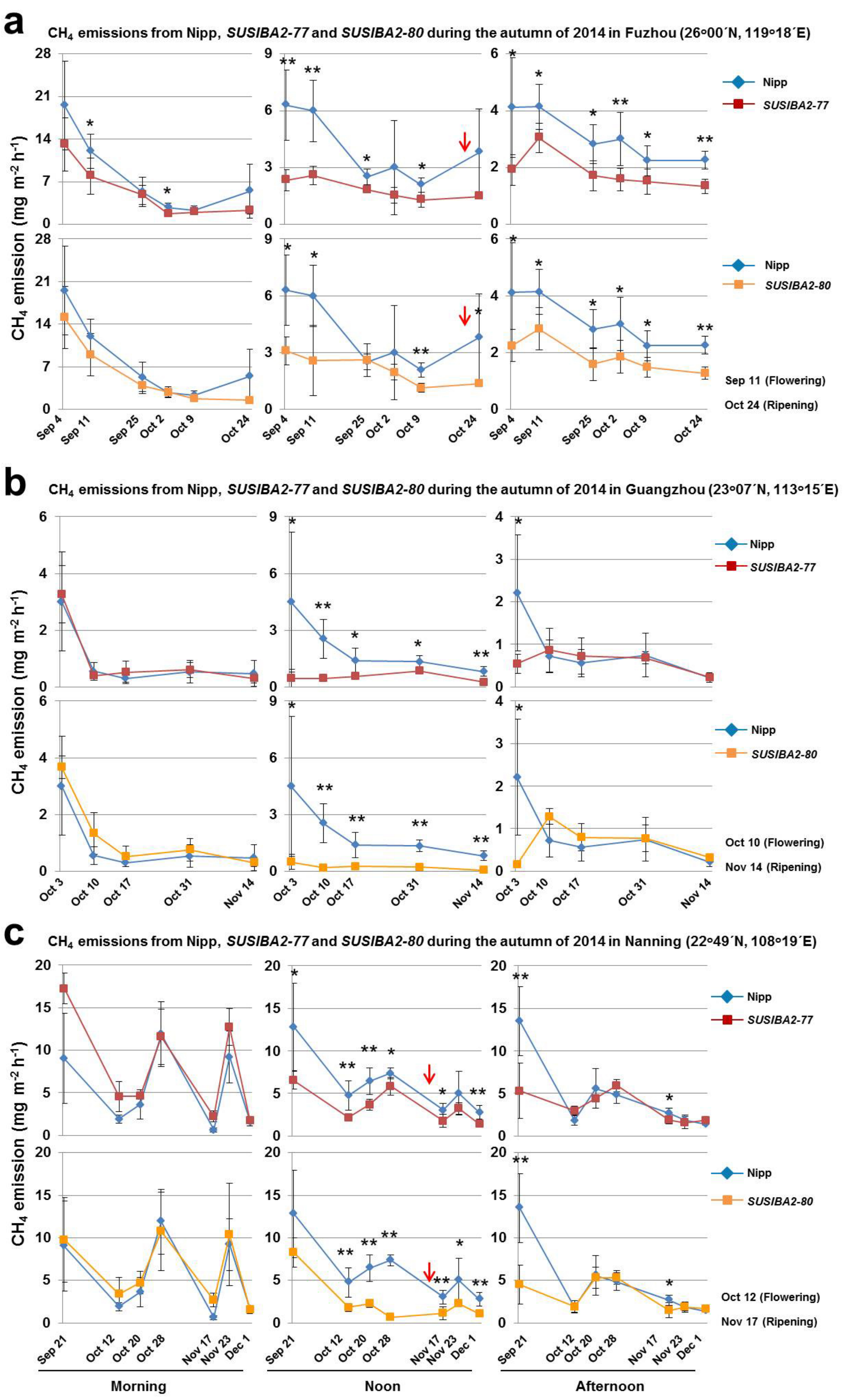
19. Nalawade, S., Nalawade, S., Liu, C., Jansson, C. & Sun, C. Development of an efficient tissue culture after crossing (TCC) system for transgenic improvement of barley as a bioenergy crop. *Appl. Energy* **91,** 405–411 (2012).
20. Sun, C., Sathish, P., Ahlandsberg, S., Deiber, A. & Jansson, C. The two genes encoding starch-branching enzymes IIa and IIb are differentially expressed in barley. *Plant Physiol.* **118,** 37–49 (1998).
21. Hiei, Y., Ohta, S., Komari, T. & Kumashiro, T. Efficient transformation of rice (*Oryza sativa* L.) mediated by *Agrobacterium* and sequence analysis of the boundaries of the T-DNA. *Plant J.* **6,** 271–282 (1994).
22. Jain, M., Nijhawan, A., Tyagi, A. K. & Khurana, J. P. Validation of housekeeping genes as internal control for studying gene expression in rice by quantitative real-time PCR. *Biochem. Biophys. Res. Commun.* **345,** 646–651 (2006).
23. Caldana, C., Scheible, W. R., Mueller-Roeber, B. & Ruzicic, S. A quantitative RT–PCR platform for high-throughput expression profiling of 2500 rice transcription factors. *Plant Methods* **3,** 7 (2007).
24. Yu, Y., Lee, C., Kim, J. & Hwang, S. Group-specific primer and probe sets to detect methanogenic communities using quantitative real-time polymerase chain reaction. *Biotechnol. Bioeng.* **89,** 670–679 (2005).
25. Westerholm, M. *et al.* Quantification of syntrophic acetate-oxidizing microbial communities in biogas processes. *Environ. Microbiol. Rep.* **3,** 500–505 (2011).
26. Narihiro, T. & Sekiguchi, Y. Oligonucleotide primers, probes and molecular methods for the environmental monitoring of methanogenic archaea. *Microb. Biotechnol.* **4,** 585–602 (2011).
27. Harrison, R. M., Yamulki, S., Goulding, K. W. T. & Webster, C. P. Effect of fertilizer application on NO and N_2O fluxes from agricultural fields. *J. Geophys. Res.* **100,** 25923–25931 (1995).
28. Westerholm, M., Hansson, M. & Schnürer, A. Improved biogas production from whole stillage by co-digestion with cattle manure. *Bioresour. Technol.* **114,** 314–319 (2012).
29. Yang, S., Peng, S., Xu, J., Luo, Y. & Li, D. Methane and nitrous oxide emissions from paddy field as affected by water-saving irrigation. *Phys. Chem. Earth* **53–54,** 30–37 (2012).



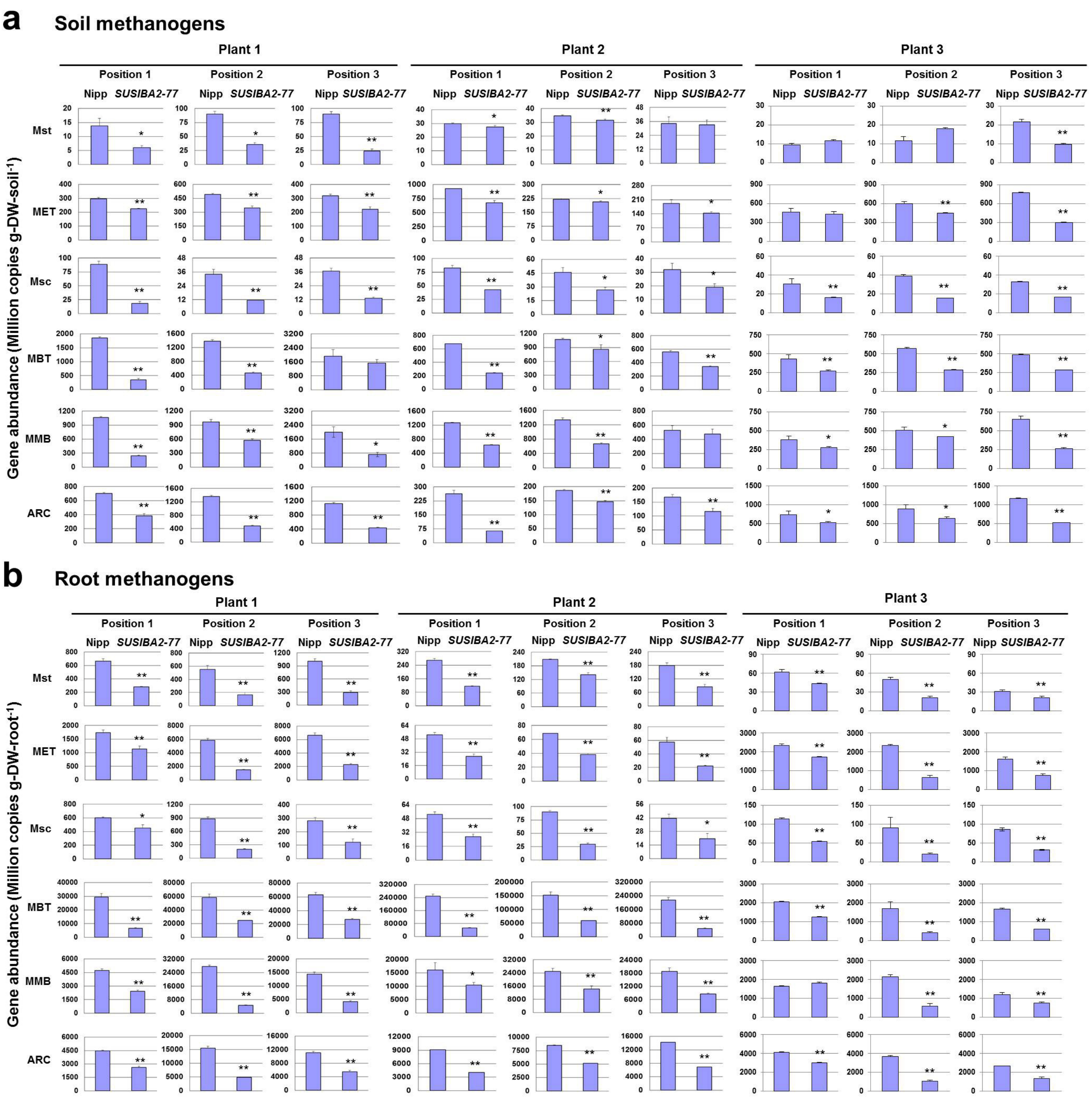
Extended Data Figure 1 | Validation of an expression cassette containing barley *SBEIIb* promoter and barley *SUSIBA2* (*HvSBEIIb* p:*HvSUSIBA2*) in a binary vector and rice genome. **a**, Construction of an expression cassette containing *HvSBEIIb* p:*HvSUSIBA2* in a binary vector was performed as described in the Methods. **b**, Validation of the construct in the rice genome was performed by PCR-based cloning using primers and HindIII-adaptor ligation, followed by sequencing. Two insertion sites were identified in rice chromosome 2, from nucleotides 57443 to 57444 and 57447 to 57448 (GenBank accession number AP006069), respectively.



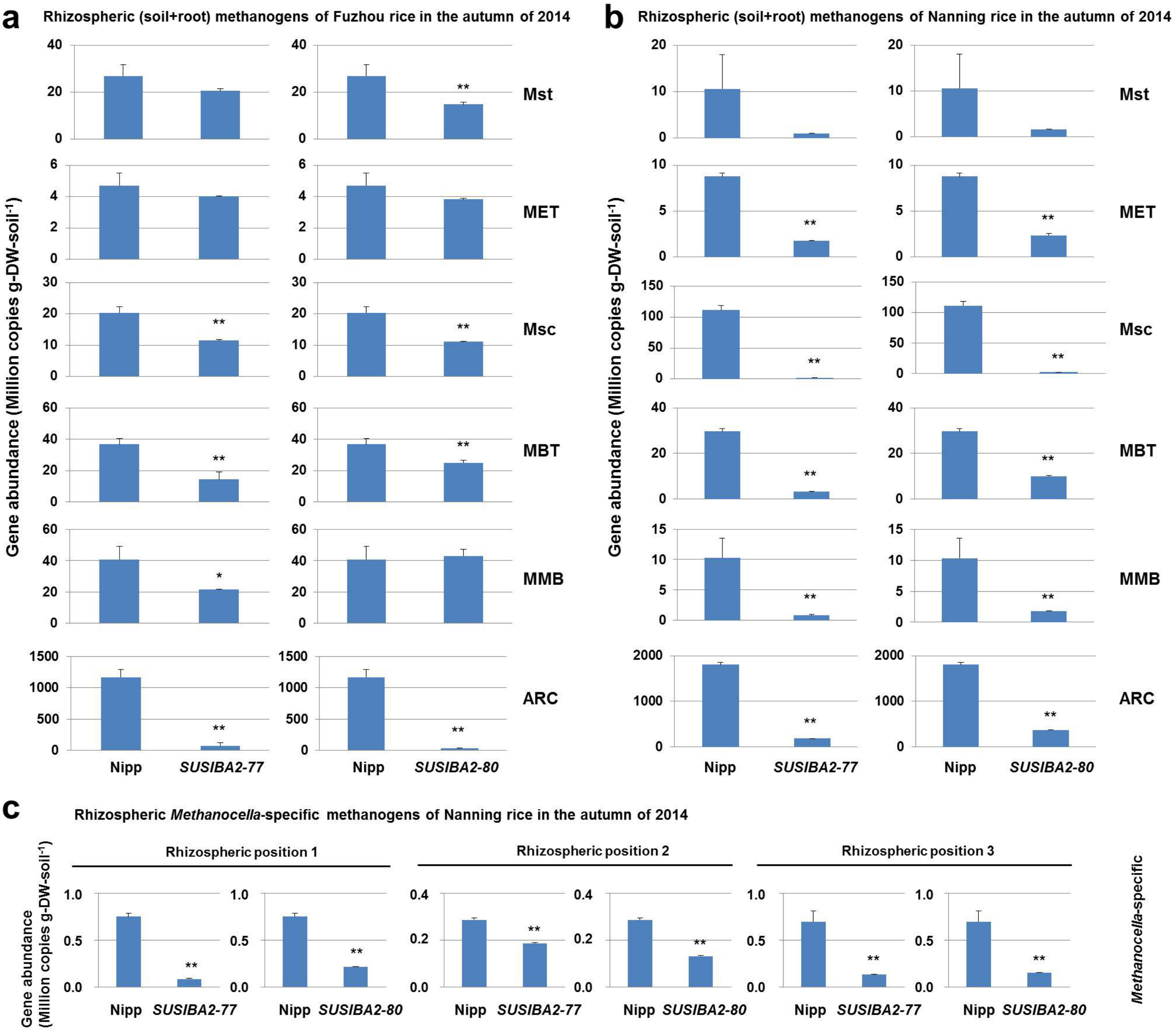
Extended Data Figure 2 | Methane emissions of *SUSIBA2* rice compared with Nipponbare (Nipp) in phytotrons. **a**, Methane emission of Nipp, *SUSIBA2-77* and *SUSIBA2-80* rice (15 daf). Six independent plants ($n = 6$) from each rice line were used for measurements. **b**, Methane emission of Nipp and *SUSIBA2-77* (28 daf) at 30 min (left panel) or 60 min (right panel) after plants were covered. A linear relationship over time for the measured methane concentrations at 30 and 60 min was found, as presented by similar methane fluxes determined from the different time points. Three plants ($n = 3$) from each rice line were used for time point measurements. A statistically significant reduction of methane emission in *SUSIBA2* rice is indicated (one-way ANOVA, $^{**}P \leq 0.01$ or $^{*}P \leq 0.05$, error bars show s.d.).


a
CH$_4$ emissions from Nipp, *SUSIBA2-77* and *SUSIBA2-80* during the autumn of 2014 in Fuzhou (26°00´N, 119°18´E)
CH$_4$ emission (mg m^{-2} h^{-1})
Nipp
SUSIBA2-77
Nipp
SUSIBA2-80
Sep 11 (Flowering)
Oct 24 (Ripening)
0
7
14
21
28
0
3
6
9
0
2
4
6
Sep 4
Sep 11
Sep 25
Oct 2
Oct 9
Oct 24
*
**
b
CH$_4$ emissions from Nipp, *SUSIBA2-77* and *SUSIBA2-80* during the autumn of 2014 in Guangzhou (23°07´N, 113°15´E)
CH$_4$ emission (mg m^{-2} h^{-1})
Nipp
SUSIBA2-77
Nipp
SUSIBA2-80
Oct 10 (Flowering)
Nov 14 (Ripening)
0
2
4
6
0
3
6
9
0
1
2
3
4
Oct 3
Oct 10
Oct 17
Oct 31
Nov 14
c
CH$_4$ emissions from Nipp, *SUSIBA2-77* and *SUSIBA2-80* during the autumn of 2014 in Nanning (22°49´N, 108°19´E)
CH$_4$ emission (mg m^{-2} h^{-1})
Nipp
SUSIBA2-77
Nipp
SUSIBA2-80
Oct 12 (Flowering)
Nov 17 (Ripening)
0
5
10
15
20
Sep 21
Oct 12
Oct 20
Oct 28
Nov 17
Nov 23
Dec 1
Morning
Noon
Afternoon

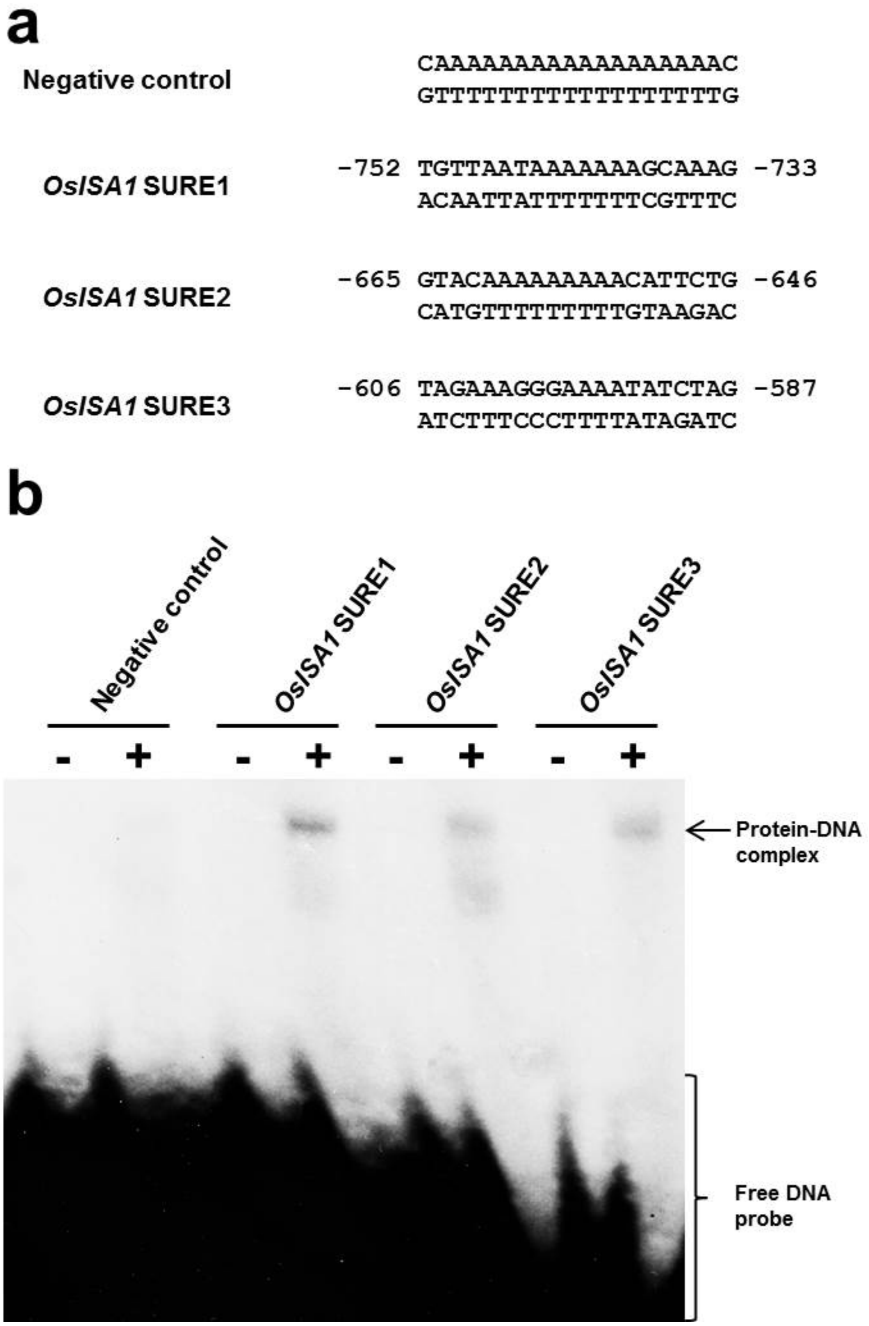
Extended Data Figure 3 | Diurnal and seasonal methane emissions from *SUSIBA2-77* and *SUSIBA-80* rice compared with Nipponbare (Nipp) in autumn 2014 at three sites in China. Methane emissions of six independent plants ($n = 6$) from three time points during the day (morning 8.00 a.m., noon 12.00 p.m. and afternoon 4.00 p.m.) on different dates are presented. Key rice development stages for the corresponding dates are indicated. Time points for sampling soil and roots for methanogen analysis are indicated by red arrows. **a**, Methane emission from Fuzhou. **b**, Methane emission from Guangzhou. **c**, Methane emission from Nanning. Reported statistically significant reduction of methane emission in *SUSIBA2* rice is indicated (one-way ANOVA, $^{**}P \leq 0.01$ or $^{*}P \leq 0.05$, error bars show s.d.).



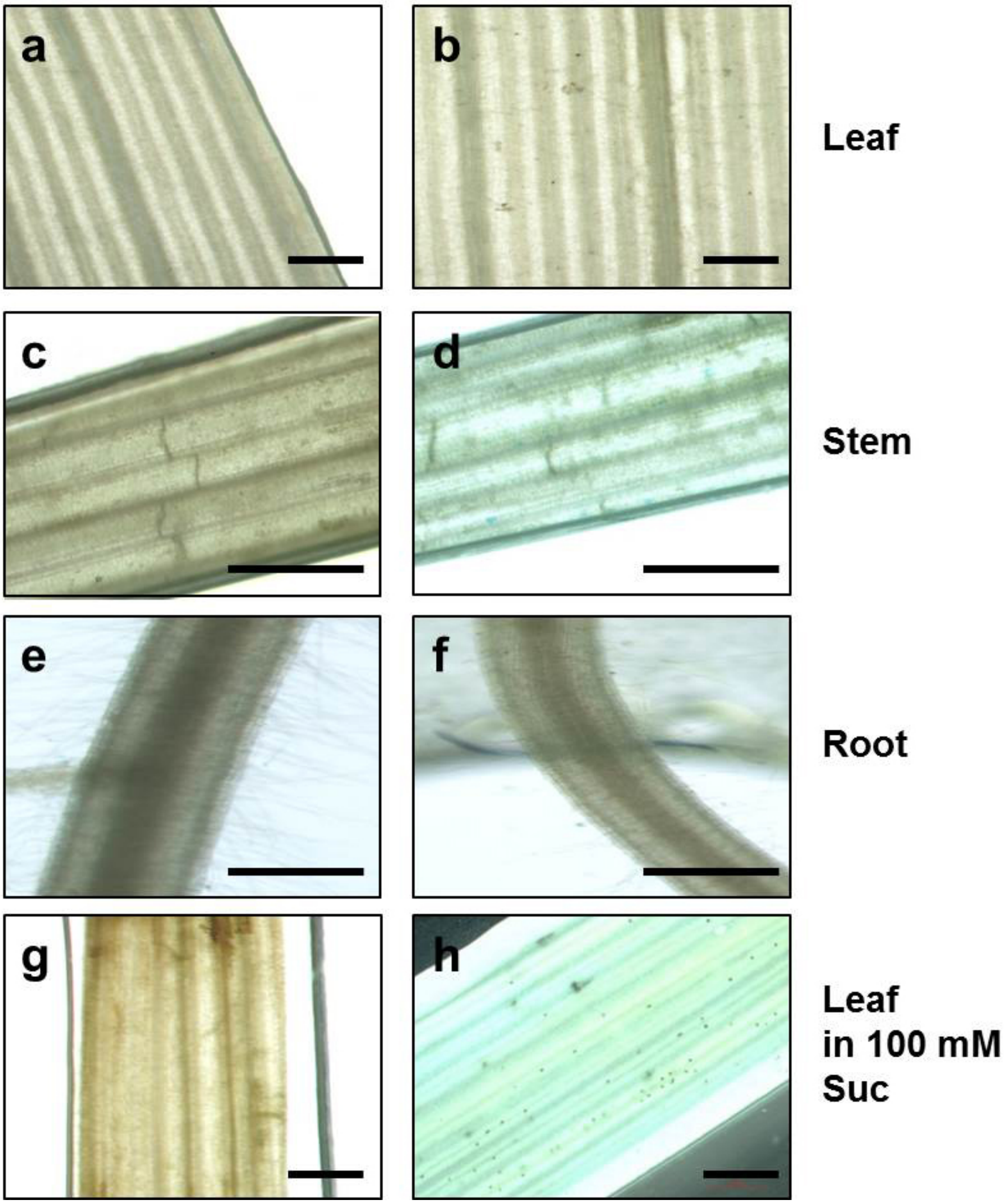
Extended Data Figure 4 | qPCR quantification of rhizospheric methanogenic communities associated with *SUSIBA2-77* rice and Nipp in phytotrons. Soil and root samples from three independent positions, positions 1–3 ($n = 3$), in the underground vicinity close to the root tip and proximal regions of three independent plants ($n = 3$) for Nipp and *SUSIBA2-77* rice were collected and analysed. Technical triplicates per position were applied. Six pairs of primers (Supplementary Table 1) were used to quantify total archaea (ARC) and methanogens (MET), and the orders Methanobacteriales (MBT), Methanomicrobiales (MMB) and Methanocellales and two families Methanosaetaceae (Mst) and Methanosarcinaceae (Msc) of the order Methanosarcinales, respectively. Results from soil and root samples are shown in **a** and **b**, respectively. Existing numbers of all methanogenic groups and total archaea were significantly reduced in the *SUSIBA2* rhizosphere compared to that of Nipp (one-way ANOVA, $^{**}P \leq 0.01$ or $^{*}P \leq 0.05$, error bars show s.d.).



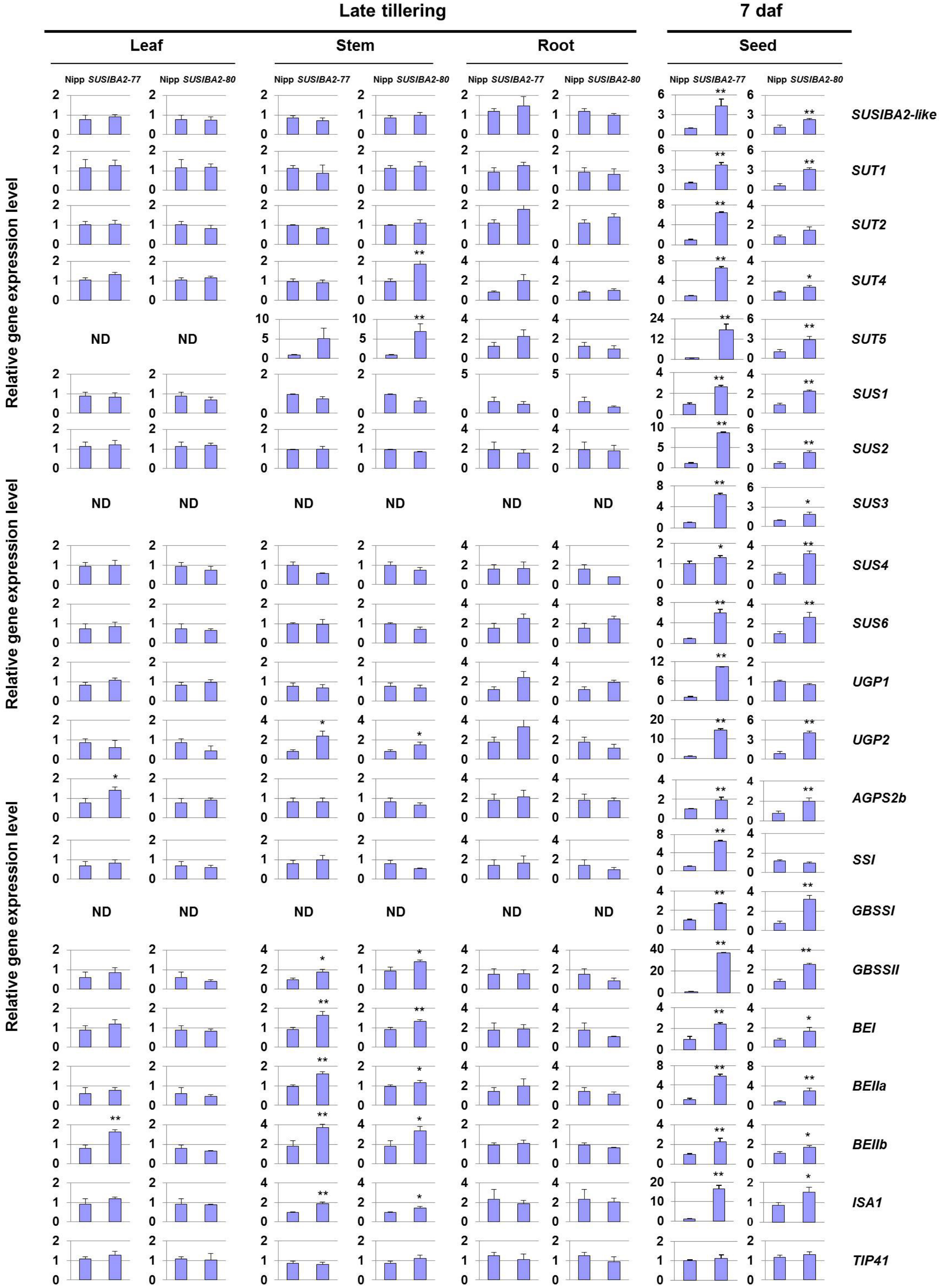
Extended Data Figure 5 | qPCR quantification of rhizospheric methanogens associated with *SUSIBA2* rice and Nipponbare (Nipp) from rice paddies. Soil and root samples (a mixture) from three positions ($n = 3$) close to the root tip and proximal regions of three independent plants ($n = 3$) for Nipp, *SUSIBA2-77* and *SUSIBA2-80* rice, respectively, were collected. The sampling time and sites are indicated in Extended Data Fig. 3. Technical triplicates per position were applied. Six pairs of primers (Supplementary Table 1) were used to quantify total archaea (ARC) and methanogens (MET), and the orders Methanobacteriales (MBT), Methanomicrobiales (MMB) and Methanocellales and two families Methanosaetaceae (Mst) and Methanosarcinaceae (Msc) of the order Methanosarcinales, respectively. Primers (Supplementary Table 1) specific to Methanocella were also used for quantification. **a**, Methanogenic communities in samples from Fuzhou. **b**, Methanogenic communities in samples from Nanning. **c**, *Methanocella* in samples from Nanning. All methanogenic groups and total archaea were significantly reduced in the *SUSIBA2* rice rhizosphere compared with Nipp (one-way ANOVA, $^{**}P \leq 0.01$ or $^{*}P \leq 0.05$, error bars show s.d.).



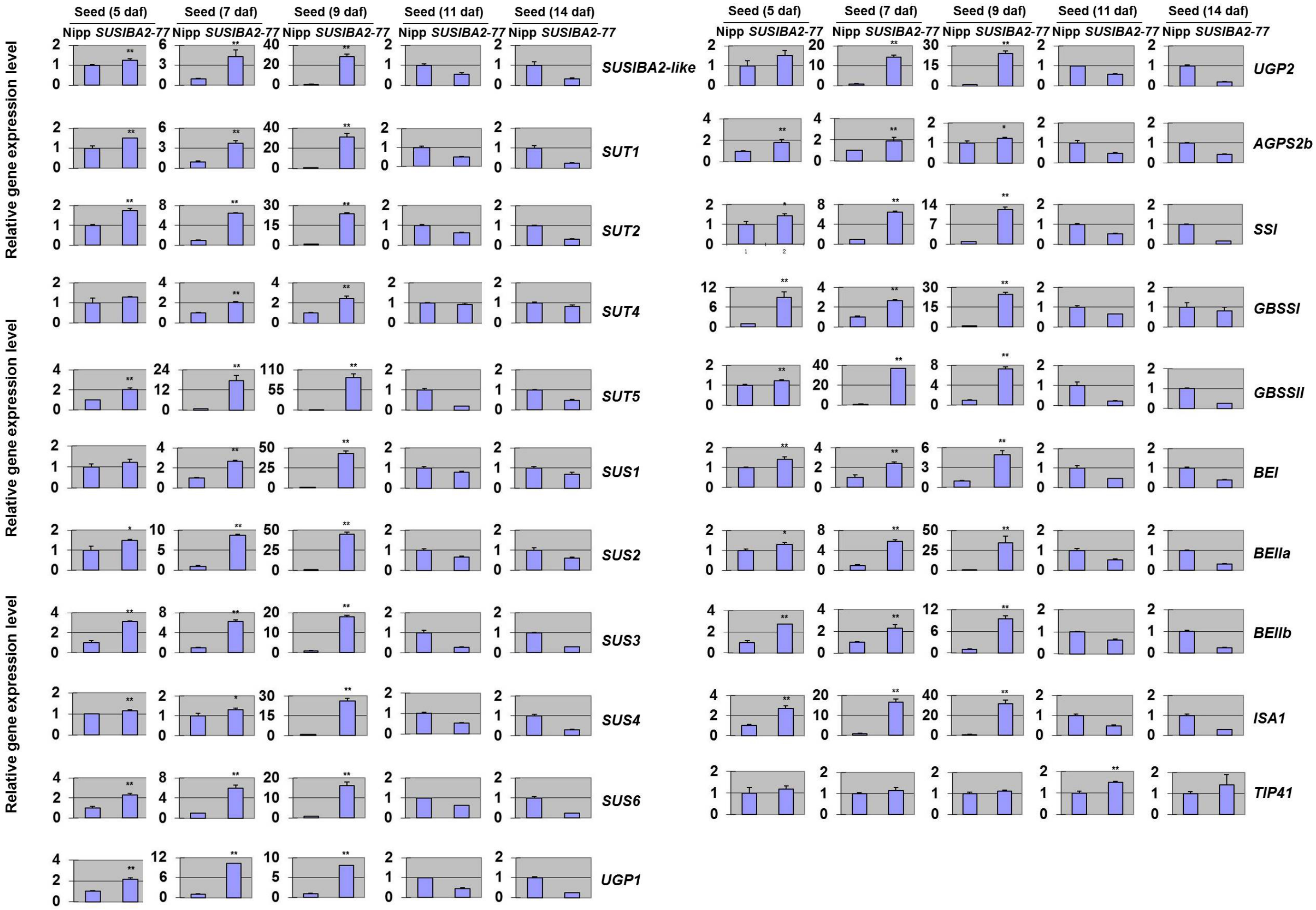
Extended Data Figure 6 | Binding activity of HvSUSIBA2 to SURE sequences in the rice *ISA1* promoter. **a**, Three SURE sequences in the rice *ISA1* promoter (GenBank accession number AB093426) were used and a negative ('A stretch') control was included. **b**, Barley SUSIBA2 protein (HvSUSIBA2) was overexpressed from *E. coli* and used for electrophoretic mobility shift assay (EMSA).



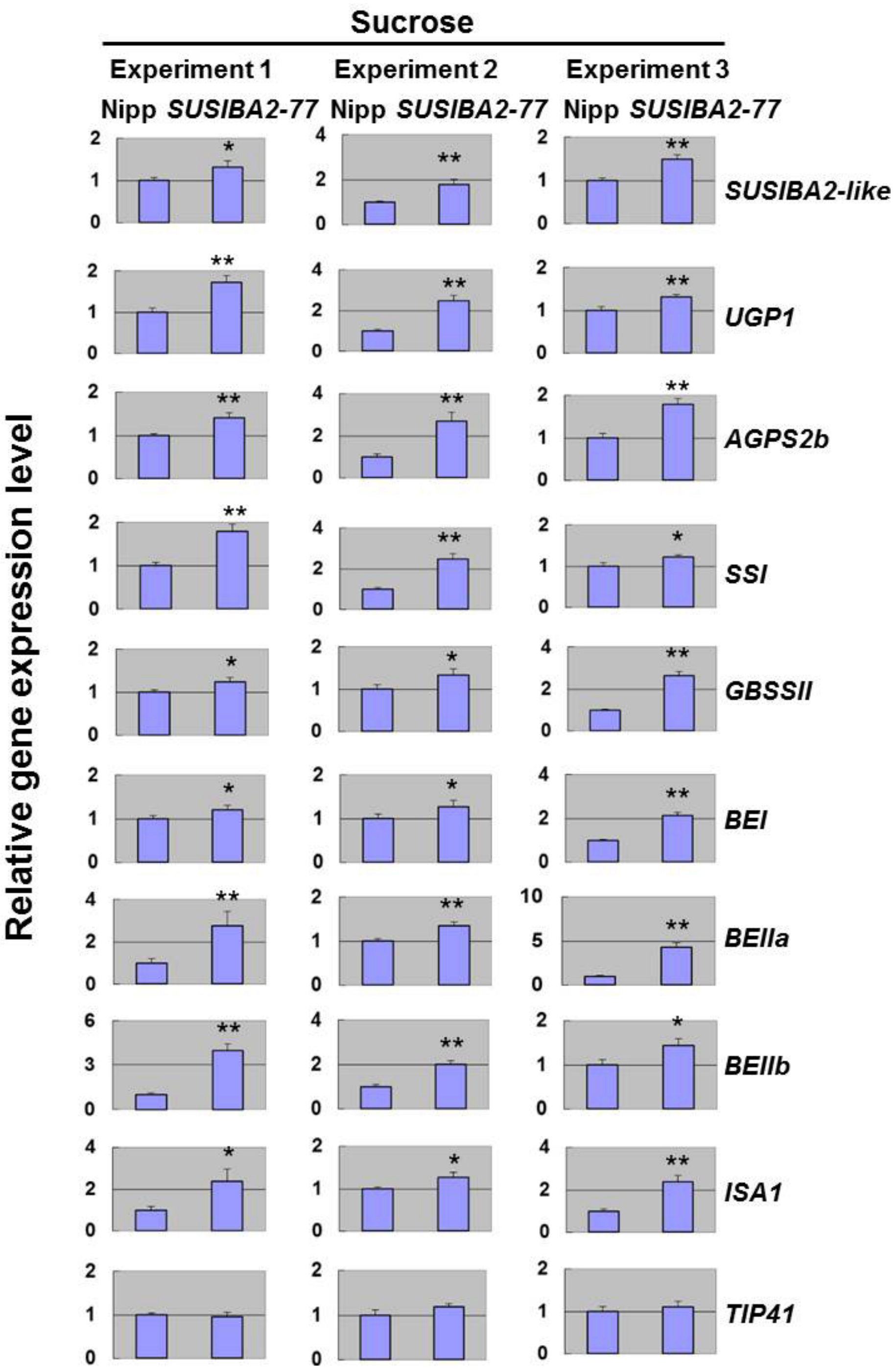
Extended Data Figure 7 | *HvSBEIIb* promoter activity analysis in rice seedlings. **a**, *HvSBEIIb* p:*GUS* was introduced in Nipponbare (Nipp). GUS activity was stained in different tissues of transformant and Nipp lines. The GUS activity was found in transformant stems and induced by sucrose (Suc) in leaves. **a**, Nipp leaf. **b**, Transformant leaf. **c**, Nipp stem. **d**, Transformant stem. **e**, Nipp root. **f**, Transformant root. **g**, Nipp leaf induced by 100 mM sucrose. **h**, Transformant leaf induced by 100 mM sucrose. Scale bars, 2 mm.


Late tillering
7 daf
Leaf
Stem
Root
Seed
Nipp SUSIBA2-77
Nipp SUSIBA2-80
Relative gene expression level
SUSIBA2-like
SUT1
SUT2
SUT4
SUT5
SUS1
SUS2
SUS3
SUS4
SUS6
UGP1
UGP2
AGPS2b
SSI
GBSSI
GBSSII
BEI
BEIIa
BEIIb
ISA1
TIP41
ND
*
**

Extended Data Figure 8 | Transcriptomic analysis of genes related to sugar metabolism in late tillering plants of *SUSIBA2* rice and Nipponbare (Nipp). Relative expression levels of 23 genes together with a housekeeping gene, *TIP41-like,* were analysed by qPCR and compared between Nipp and *SUSIBA2-77* and *SUSIBA2-80* rice in leaves, stems and roots at late tillering stage and in seeds at 7 daf. Three plants were used ($n = 3$) and technical triplicates were performed for each rice line. *SUT3* and *SUS5/7* transcripts were not detected in all samples (not shown) and *HvSUSIBA2* expression was similar in *SUSIBA2-77* and *SUSIBA2-80* rice, as presented in Fig. 3b. One-way ANOVA was used for statistical analysis (*$P \leq 0.05$ or **$P \leq 0.01$, error bars show s.d.). ND, not detected.



Extended Data Figure 9 | Transcriptomic analysis of 23 genes related to sugar metabolism in developing rice seeds of *SUSIBA2* rice and Nipponbare (Nipp). Relative expression levels of 23 genes together with a housekeeping gene, *TIP41-like,* were analysed by qPCR and compared between *SUSIBA2-77* rice and Nipp at 5, 7, 9, 11 and 14 daf. Seeds from three independent plants were used ($n = 3$) and technical triplicates were performed. *SUT3* and *SUS5/7* transcripts were not detected in all samples (not shown) and *HvSUSIBA2* expression is presented in Fig. 3b. One-way ANOVA was used for statistical analysis (*$P \leq 0.05$ or **$P \leq 0.01$, error bars show s.d.).



Extended Data Figure 10 | Sugar-induction in Nipponbare (Nipp) and *SUSIBA2* rice. Rice leaves of Nipp and *SUSIBA2-77* were depleted for sucrose in the dark before induction with 100 mM sucrose. Sugar induction was carried out in the dark for 24 h. An enhanced expression of HvSUSIBA2-regulated genes (9 of the 12 genes except *SUT5*, *SUS1* and *GBSSI* that were not detected in either rice cultivar under dark conditions) in *SUSIBA2-77* was observed compared with Nipp. One-way ANOVA was used for statistical analysis (*$P \leq 0.05$ or **$P \leq 0.01$, error bars show s.d.). The experiment was repeated at least three times ($n = 3$).

doi:10.1038/nature14650

Lanosterol reverses protein aggregation in cataracts

Ling Zhao[1,2,3]*†, Xiang-Jun Chen[4]*, Jie Zhu[3,5]*, Yi-Bo Xi[4]*, Xu Yang[6]*, Li-Dan Hu[4]*, Hong Ouyang[2,3], Sherrina H. Patel[3], Xin Jin[6], Danni Lin[3], Frances Wu[3], Ken Flagg[3], Huimin Cai[1,7], Gen Li[1], Guiqun Cao[1], Ying Lin[2,3], Daniel Chen[3], Cindy Wen[3], Christopher Chung[3], Yandong Wang[2], Austin Qiu[3,8], Emily Yeh[3], Wenqiu Wang[3,9], Xun Hu[1], Seanna Grob[3], Ruben Abagyan[10], Zhiguang Su[1], Harry Christianto Tjondro[4], Xi-Juan Zhao[4], Hongrong Luo[3], Rui Hou[7], J. Jefferson P. Perry[11], Weiwei Gao[3,12], Igor Kozak[13], David Granet[3], Yingrui Li[6], Xiaodong Sun[9], Jun Wang[6], Liangfang Zhang[3,12], Yizhi Liu[2], Yong-Bin Yan[5] & Kang Zhang[1,2,3,12,14]

The human lens is comprised largely of crystallin proteins assembled into a highly ordered, interactive macro-structure essential for lens transparency and refractive index. Any disruption of intra- or inter-protein interactions will alter this delicate structure, exposing hydrophobic surfaces, with consequent protein aggregation and cataract formation. Cataracts are the most common cause of blindness worldwide, affecting tens of millions of people[1], and currently the only treatment is surgical removal of cataractous lenses. The precise mechanisms by which lens proteins both prevent aggregation and maintain lens transparency are largely unknown. Lanosterol is an amphipathic molecule enriched in the lens. It is synthesized by lanosterol synthase (LSS) in a key cyclization reaction of a cholesterol synthesis pathway. Here we identify two distinct homozygous *LSS* missense mutations (W581R and G588S) in two families with extensive congenital cataracts. Both of these mutations affect highly conserved amino acid residues and impair key catalytic functions of LSS. Engineered expression of wild-type, but not mutant, *LSS* prevents intracellular protein aggregation of various cataract-causing mutant crystallins. Treatment by lanosterol, but not cholesterol, significantly decreased preformed protein aggregates both *in vitro* and in cell-transfection experiments. We further show that lanosterol treatment could reduce cataract severity and increase transparency in dissected rabbit cataractous lenses *in vitro* and cataract severity *in vivo* in dogs. Our study identifies lanosterol as a key molecule in the prevention of lens protein aggregation and points to a novel strategy for cataract prevention and treatment.

Cataracts account for over half of all cases of blindness worldwide, with the only established treatment involving surgical removal of the opacified lens. In developed nations, cataract surgeries amount to a significant portion of healthcare costs owing to the sheer prevalence of the disease among ageing populations. In addition, there is major morbidity associated with cataracts in developing countries, where there is limited access to surgical care.

High concentrations of crystallin proteins in lens fibres contribute to lens transparency and refractive properties[2]. The crystallin superfamily is composed of α-, β- and γ-crystallins, which are some of the most highly concentrated intracellular proteins in the human body. Protein aggregation is the single most important factor in cataract formation[3]. Factors that lead to protein aggregation include mutations in crystallin proteins, which are known to cause congenital cataracts, or oxidative stress, which in turn contributes to age-related cataracts. However, the precise mechanisms by which lens proteins maintain transparency or cause opacification are not completely understood.

Lanosterol synthase (2,3-oxidosqualene-lanosterol cyclase, LSS; EC 5.4.99.7) is encoded by the *LSS* gene. The LSS protein catalyses the conversion of (*S*)-2,3-oxidosqualene to lanosterol, which is a key early rate-limiting step in the biosynthesis of cholesterol, steroid hormones, and vitamin D (ref. 4). LSS was found to be expressed in the lens[5]. It was previously reported that the specific combination of hypomorphic mutations on *LSS* and *FDFT1* (farnesyl diphosphate farnesyl transferase 1) could decrease cholesterol levels in the lens and result in cataracts in Shumiya cataract rats (SCR)[6]. Here we identify novel homozygous mutations in the *LSS* gene in two consanguineous families and investigate the ability of lanosterol to alleviate protein aggregation and cataract formation.

We identified three children with severe congenital cataract from a consanguineous family of Caucasian descent (Fig. 1a). We performed whole-exome sequencing to an average of no less than 55-fold depth coverage on the target region (Extended Data Table 1a) in order to identify the causal mutation. On average, ~60,800–80,800 SNPs were detected in each exome (Extended Data Table 1b). Using a consanguineous recessive model and filtering against common variants (minor allele frequency >0.5%) in public databases, including dbSNP and the 1000 Genomes Project, as well as mutation function predictions (predicted by SIFT[7], Polyphen2[8], Phylop[9] and Mutationtaster[10]), we narrowed down potential candidate gene variants and identified a variant (G588S) in *LSS* on chromosome 21 as the most likely candidate (Extended Data Table 1c). Three affected children were homozygous for the G→A transition (G588S) in *LSS*, (GRch37/hg19: chr21:47615645; NM_001001438.2:c.1762G > A, NM_001001438.1: p.G588S), while the unaffected father, mother and remaining child were heterozygous for the change (Fig. 1a, b). Whole-genome SNP genotyping identified three long continuous homozygous regions in this family by HomozygosityMapper[11] (chr2:q22.1–q24.1, chr2:q31.1–q32.1 and chr21:q22.3; Extended Data Fig. 1a and Extended Data Table 1d). The *LSS* gene was located in one of the homozygous regions on chromosome 21 (Extended Data Fig. 1b). Furthermore, we screened for mutations in the *LSS* gene in 154 families with congenital cataracts and identified another homozygous mutation, W581R (GRch37/hg19: chr21:47615666; NM_001001438.2:c.1741T > C, NM_001001438.1: p.W581R), in a second consanguineous family (Fig. 1a, b, c). These two mutations were absent in 11,000 control chromosomes.

The amino acid residues W581 and G588 in LSS are highly conserved (Fig. 2a). We performed computational modelling analysis to

[1]Molecular Medicine Research Center, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu 610041, China. [2]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China. [3]Department of Ophthalmology and Biomaterials and Tissue Engineering Center, Institute for Engineering in Medicine, University of California San Diego, La Jolla, California 92093, USA. [4]State Key Laboratory of Membrane Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China. [5]Department of Ophthalmology, Xijing Hospital, Fourth Military Medical University, Xi'an 710032, China. [6]BGI-Shenzhen, Shenzhen 518083, China. [7]Guangzhou KangRui Biological Pharmaceutical Technology Company, Guangzhou 510005, China. [8]CapitalBio Genomics Co., Ltd., Dongguan 523808, China. [9]Department of Ophthalmology, Shanghai First People's Hospital, School of Medicine, Shanghai JiaoTong University, Shanghai 20080, China. [10]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, California 92093, USA. [11]Department of Biochemistry, University of California Riverside, Riverside, California 92521, USA. [12]Department of Nanoengineering, University of California, San Diego, La Jolla, California 92093, USA. [13]King Khaled Eye Specialist Hospital, Riyadh, Kingdom of Saudi Arabia. [14]Veterans Administration Healthcare System, San Diego, California 92093, USA. †Present address: Institute of Molecular Medicine, Peking University, Beijing 100871, China.
*These authors contributed equally to this work.

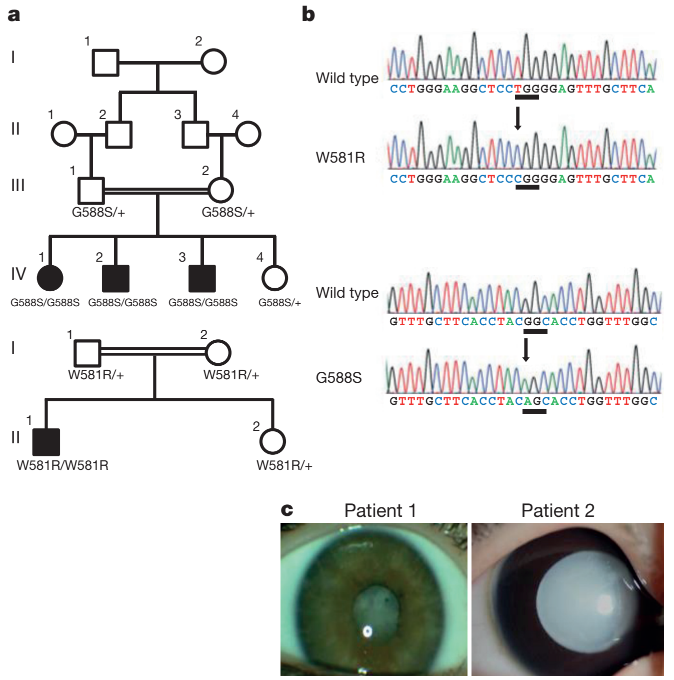


Figure 1 | Identification of mutations in *LSS* causing congenital cataracts. **a**, Pedigrees of affected families and cataract phenotype. Squares and circles indicate males and females respectively. +, wild-type allele; W581R and G588S are the two mutations. **b**, Upper panel, DNA sequencing data of an unaffected individual and an affected child (II-1) with a homozygous W581R mutation; lower panel, DNA sequencing data of an unaffected individual and an affected child (IV-1) with a homozygous G588S mutation. The underlined sequence indicates the nucleic acid change. **c**, Left, colour photograph of patient 1's right eye in the first pedigree (IV-1) with a total cataract; right, colour photograph of patient 2's right eye in the same pedigree (IV-3) with a cataract.

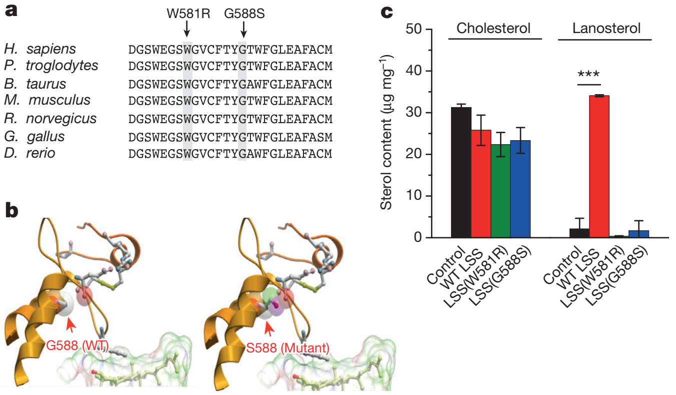


Figure 2 | LSS mutations abolished the cyclase enzymatic function. **a**, Conservation of W581R and G588 in LSS across several species: *Homo sapiens*, *Pan troglodytes*, *Bos taurus*, *Mus musculus*, *Rattus norvegicus*, *Gallus gallus* and *Danio rerio*. **b**, Computer modelling of LSS structure and impact of the LSS W581R and G588S mutations. A computer modelling analysis identifies a loop originating from C584 and ending at E578 with the key side chain of W581 at the tip of the loop stabilizing the sterol. The loop is fixed by an S–S bridge and the E578–R639 salt bridge. Amide nitrogen N of G588 interacts with the C584 from the previous helical turn and the Cα hydrogen of G588 is in close proximity to the critical E578, which then forms a strong salt bridge with R639 of the same supporting helix. The mutation G588S causes the side chain of the serine to clash into the E578 residue of the loop and is incompatible with the structure. Arrow indicates the location of the mutant side chain. **c**, Effect of engineered expression of the wild-type protein (WT LSS) and LSS mutants on sterol content. Wild-type LSS markedly increased lanosterol production, whereas neither W581R nor the G588S mutant exhibited any cyclase activity. $n = 3$ in each group; $^{***}P < 0.001$.

investigate the effects of the W581R and G588S mutations on the 3D structure and function of LSS. The amino acid tryptophan at position 581 has been reported to contribute to the catalytic site of the cyclase activity[12]. The G588S mutant was modelled by in-place replacement followed by side-chain refinement. The S588 side-chain refinement could not resolve the van der Waals clash between the serine side chain and the backbone carbonyl of E578, which forms a key salt bridge with R639. The orientation of the E579:C584 loop needed to be distorted to accommodate the mutation. The side chain of the mutant S588 clashed into an adjacent loop, indicating that the mutation was incompatible with the normal enzymatic structure and function of LSS (Fig. 2b). Supporting the *in silico* results, expression of wild-type LSS in a cell-transfection experiment exhibited cyclase activity and dramatically increased the amount of lanosterol production in the lipid fraction in HeLa cells, while neither the W581R nor the G588S mutant protein demonstrated any cyclase activity (Fig. 2c).

In contrast, the cholesterol level was unaffected by the expression of wild-type or mutant LSS, suggesting that there may be an alternative pathway for cholesterol homeostasis. The W581R and G588S mutations did not alter subcellular localization or cause aggregates of LSS protein when compared to that of wild-type LSS, suggesting that the cataract phenotype was not due to the formation of light-scattering particles by mutant LSS proteins themselves (Extended Data Fig. 2).

The aggregation of crystallins, the major structural proteins in the lens, is a predominant cause of various types of cataracts[3]. To mimic protein aggregation in the cataractous lens, six known cataract-causing mutant crystallin proteins were expressed in human lens progenitor cells, human lens epithelial line B-3 (HLEB-3), or HeLa cells. These mutant crystallins formed p62-positive inclusion bodies/aggresomes in all three transfected cell lines, suggesting that aggregation is an intrinsic property of mutant crystallins (Fig. 3a and Extended Data Figs 3 and 4)[13]. Co-expression of wild-type LSS and a cataract-causing mutant crystallin protein significantly reduced both the number and size of intracellular crystallin aggregates, whereas LSS mutants failed to do so alone (Fig. 3b, c and Extended Data Figs 3 and 4). Western blot analysis indicated that the Y118D mutant of αA-crystallin was released from intracellular aggregates and became more soluble with wild-type LSS (Fig. 3d and Extended Data Fig. 4c). Furthermore, addition of lanosterol, but not cholesterol, in the culture medium of cells co-expressing an LSS mutant and a mutant crystallin successfully reduced crystallin aggregation (Fig. 3c and Extended Data Figs 3 and 4). This result indicated that lanosterol, but not cholesterol, could be an effective agent to release mutant crystallin proteins from aggregation. Supporting this hypothesis, lanosterol significantly inhibited aggresome formation of both wild-type and mutated crystallin proteins in a concentration-dependent manner, while cholesterol had no effect (Fig. 3e, f and Extended Data Fig. 5). We further showed that lanosterol, but not cholesterol, increased the amounts of mutant crystallins in the soluble fractions of cell lysates (Fig. 3g and Extended Data Fig. 6a). Using serial live-cell imaging of cells expressing a GFP-fused Y118D mutant of αA-crystallin, we showed that addition of lanosterol could effectively diminish crystallin aggregates with a half-life of 222 ± 8 min (Fig. 3h), whereas addition of DMSO or cholesterol did not reduce aggresome formation (Extended Data Fig. 6b). Single-particle tracking in live cells clearly showed that lanosterol has an important role in the dissociation of pre-formed intracellular protein aggregates.

To investigate whether lanosterol has a direct effect on the dissolution of aggregated proteins, the aggregates of five wild-type and nine mutant crystallins were obtained by heating wild-type and mutated crystallins in the presence of 1 M guanidine chloride. Under this condition, all crystallin proteins formed amyloid-like fibrils as revealed by the enhancement of thioflavin T (ThT) fluorescence, the fibrillar structures under negatively stained transmission electron microscopy (TEM), and the low turbidity value (Fig. 4 and Extended Data Fig. 6c).

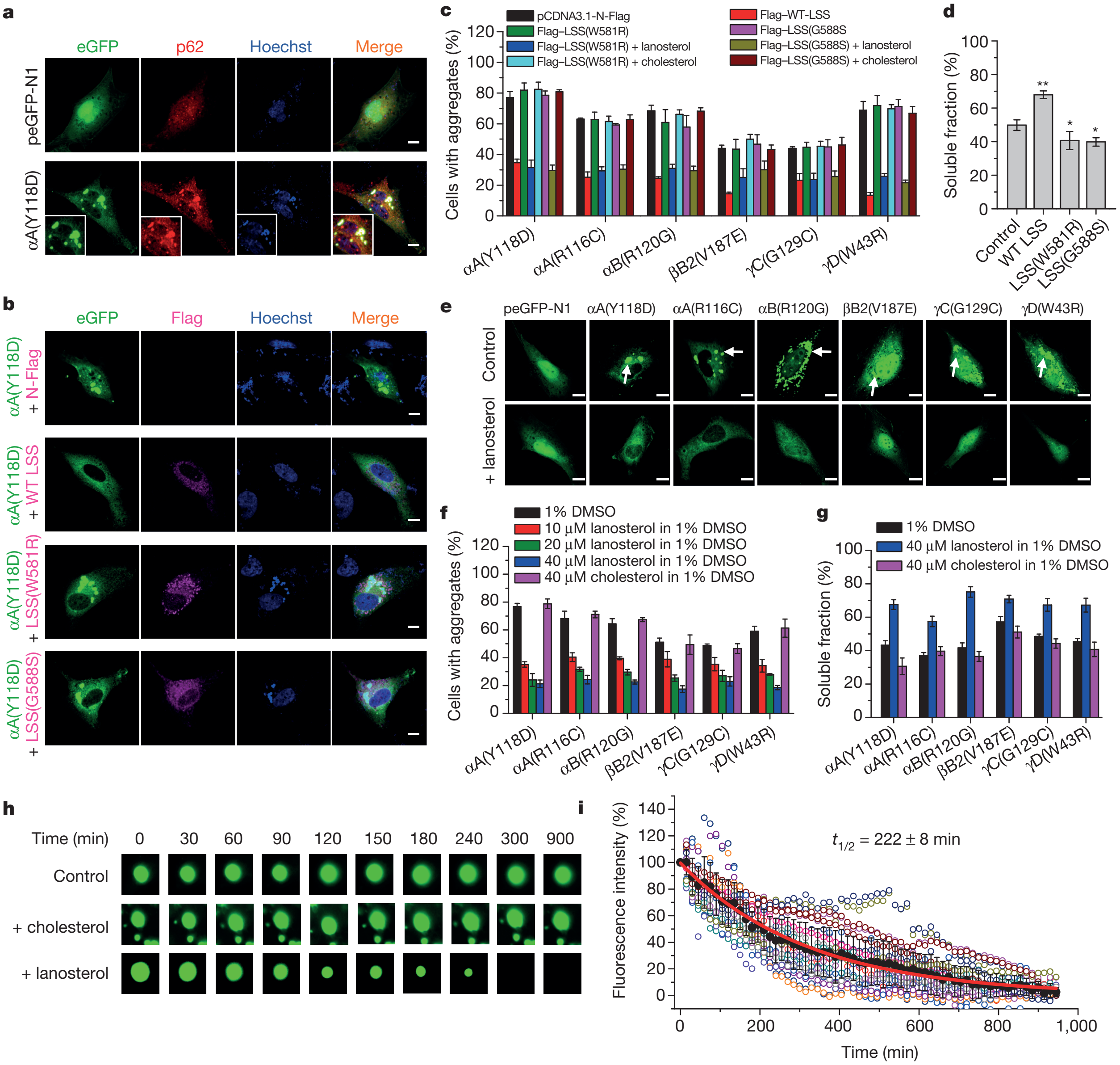


Figure 3 | Lanosterol reduced intracellular aggregation of various crystallin mutant proteins. a, Confocal images of crystallin protein aggregates in human lens progenitor cells. The cataract-causing Y118D mutant of αA-crystallin formed p62-positive intracellular inclusion bodies or aggresomes. Green, eGFP–crystallin proteins; red, p62; blue, nuclei. Cells transfected with peGFP-N1 were used as a control. **b**, Confocal images of the inhibitory effect of LSS on crystallin aggregates. **c**, Inhibition of crystallin mutant aggregation by wild-type LSS (WT LSS) and lanosterol, but not mutant LSS or cholesterol. **d**, Increase in soluble αA-crystallin(Y118D) mutant protein by co-expression of wild-type LSS but not LSS mutants (Y118D co-expressed with pcDNA3.1–N-Flag was used as a control). Quantitative analysis was performed using densitometry of crystallin proteins by western blot analysis of the supernatant or insoluble fraction of cell lysates. $n = 3$ in each group; representative western blot analysis is shown in Extended Data Fig. 4c; $*P < 0.05$, $**P < 0.01$. **e**, Confocal images of the re-dissolution of preformed crystallin aggregates by lanosterol. Arrows indicate the presence of crystallin aggregation. **f**, Lanosterol significantly reduced the intracellular aggregation by various cataract-causing mutant crystallin proteins in a concentration-dependent manner. $n = 3$; $P < 1 \times 10^{-4}$. Cholesterol did not reduce intracellular aggregation. $n = 3$; $P > 0.1$. **g**, Lanosterol increased the soluble fractions of various crystallin mutants in human lens progenitor cells. $n = 3$; $P < 0.001$. **h**, Effects of DMSO, cholesterol or lanosterol on αA-crystallin(Y118D) aggregates in human lens progenitor cells by serial live-cell imaging. Progression of crystallin aggregation dissolution by lanosterol can be observed, as evidenced by decreased green fluorescence following the time-course. **i**, Effect of lanosterol on dissolution of intracellular crystallin aggregates over time. $n = 22$ from three biological replicates. The 22 repetitions are shown in open circles distinguished by different colours. The mean ± s.d. values are shown as filled black circles and error bars. The data are best fitted by the single exponential decay process (red line). Scale bars, 10 μm.

The morphology of the amyloid-like fibrils obtained here was similar to those crystallin proteins reported previously[14]. We used PBS-containing liposomes formed by dipalmitoyl phosphatidylcholine (DPPC) to increase the solubility of sterol compounds and mimic the condition of sterols in cell membranes. Lanosterol, but not cholesterol, successfully re-dissolved the aggregated crystallin proteins from the amyloid-like fibrils in a concentration-dependent manner as indicated by the disappearance of fibrillar structures in the negatively stained TEM photographs and the decrease in ThT fluorescence intensity (Fig. 4 and Extended Data Fig. 6d). As an example, the re-dissolved αA-crystallins could be identified in negatively stained TEM pictures and were around 15 nm in size (Fig. 4a)[15].

To assess the effect of lanosterol on cataract reduction in lens tissues, we isolated naturally occurring cataractous lenses from rabbits. We incubated these cataractous lenses in a 25 mM lanosterol solution for 6 days and compared lens clarity before and after treatment of

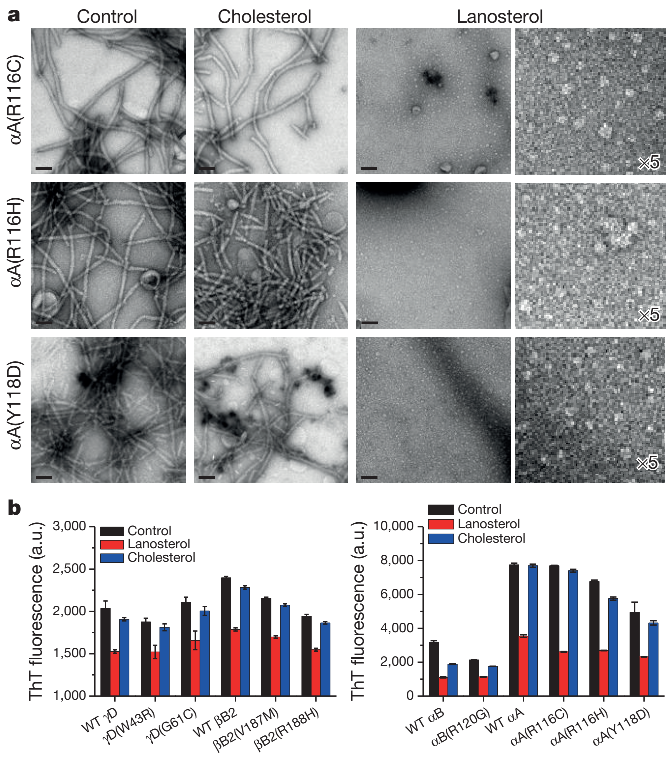


Figure 4 | Lanosterol re-dissolved pre-formed amyloid-like fibrils of crystallin proteins. a, Negatively stained TEM photographs of aggregates of αA-crystallin mutant proteins treated by a liposome vehicle, cholesterol or lanosterol in liposomes. Images in the right column of the lanosterol group show a 5× magnification of the image on their left. **b**, Effect of lanosterol on the re-dissolution of crystallin aggregates by ThT fluorescence ($n = 3$). Left, β/γ-crystallin mutants; right, α-crystallin mutants. Each bar results from three independent samples.

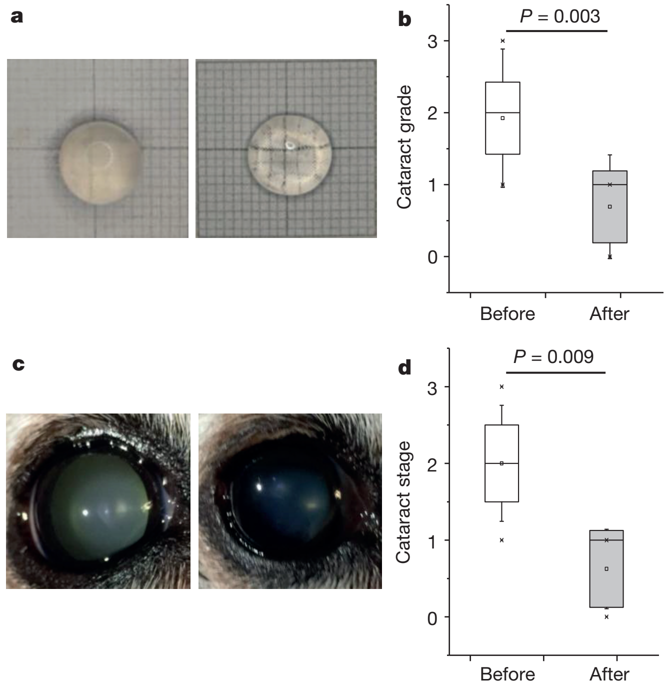


Figure 5 | Lanosterol reduced cataract severity and increased clarity. a, Photographs of a cataractous rabbit lens treated with lanosterol showing increased lens clarity. Left, before treatment; right, after. **b**, Boxplot of the quantification of the treatment effect of lanosterol ($n = 13$). **c**, Photographs of a cataractous dog lens treated with lanosterol showing increased lens clarity. Left, before treatment; right, after. **d**, Boxplot of the quantification of the treatment effect of lanosterol ($n = 7$). Range, median (horizontal line) and mean (circle) are presented. Crosses indicate the maximum and minimum cataract grades measured. Whiskers indicate the standard deviation and the box encompasses a 40% confidence interval.

lanosterol. We observed a strong trend of reduction in cataract severity, as demonstrated by an increase in lens clarity ($P < 0.003$, Wilcoxon Test, Fig. 5a, b, Extended Data Table 2a and Extended Data Fig. 7a, b). We further investigated the effect of lanosterol in reversing cataracts in dogs *in vivo*. Lanosterol treatment significantly reduced cataract severity and increased lens clarity ($P < 0.009$, Wilcoxon Test, Fig. 5c, d; Extended Data Table 2b and Extended Data Fig. 7c).

In this study, we demonstrated that homozygous mutations affecting the catalytic function of LSS cause extensive congenital cataracts with severe vision loss. The critical role of lanosterol in cataract prevention is supported by the observation that a rat strain harbouring compound *LSS* mutations recapitulates the human cataract disease phenotype[6]. Consistent with this notion, inhibition of LSS by U18666A, an LSS inhibitor (also known as an oxidosqualene cyclase inhibitor), was found to cause cataracts[16]. Furthermore, lanosterol treatment both decreased protein aggregation caused by mutant crystallin proteins in cell culture and reduced preformed cataract severity by increasing lens clarity in animal models. It is conceivable that the amphipathic nature of lanosterol allows it to intercalate into and coat hydrophobic core areas of large protein aggregates, effectively allowing these aggregations to gradually become water soluble again.

In summary, we show that lanosterol plays a key role in inhibiting lens protein aggregation and reducing cataract formation, suggesting a novel strategy for the prevention and treatment of cataracts. Cataracts are the leading cause of blindness and millions of patients every year undergo cataract surgery to remove the opacified lenses. The surgery, although very successful, is nonetheless associated with complications and morbidities. Therefore, pharmacological treatment to reverse cataracts could have large health and economic impacts. In addition, our results may have broader implications for the treatment of protein-aggregation diseases, including neurodegenerative diseases and diabetes, which collectively are a significant cause of morbidity and mortality in the elderly population, by encouraging the investigation of small-molecule approaches, such as the one demonstrated here.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 July 2014; accepted 11 June 2015.

Published online 22 July 2015.

1. Pascolini, D. & Mariotti, S. P. Global estimates of visual impairment: 2010. *Br. J. Ophthalmol.* **96,** 614–618 (2012).
2. Bloemendal, H. *et al.* Ageing and vision: structure, stability and function of lens crystallins. *Prog. Biophys. Mol. Biol.* **86,** 407–485 (2004).
3. Moreau, K. L. & King, J. A. Protein misfolding and aggregation in cataract disease and prospects for prevention. *Trends Mol. Med.* **18,** 273–282 (2012).
4. Huff, M. W. & Telford, D. E. Lord of the rings–the mechanism for oxidosqualene:lanosterol cyclase becomes crystal clear. *Trends Pharmacol. Sci.* **26,** 335–340 (2005).
5. Diehn, J. J., Diehn, M., Marmor, M. F. & Brown, P. O. Differential gene expression in anatomical compartments of the human eye. *Genome Biol.* **6,** R74 (2005).
6. Mori, M. *et al.* Lanosterol synthase mutations cause cholesterol deficiency-associated cataracts in the Shumiya cataract rat. *J. Clin. Invest.* **116,** 395–404 (2006).
7. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11,** 863–874 (2001).
8. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7,** 248–249 (2010).
9. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20,** 110–121 (2010).
10. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nature Methods* **11,** 361–362 (2014).
11. Seelow, D., Schuelke, M., Hildebrandt, F. & Nurnberg, P. HomozygosityMapper–an interactive approach to homozygosity mapping. *Nucleic Acids Res.* **37,** W593–W599 (2009).
12. Thoma, R. *et al.* Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. *Nature* **432,** 118–122 (2004).

13. Dobson, C. M. Protein folding and misfolding. *Nature* **426,** 884–890 (2003).
14. Ecroyd, H. & Carver, J. A. Crystallin proteins and amyloid fibrils. *Cell. Mol. Life Sci.* **66,** 62–81 (2009).
15. Braun, N. *et al.* Multiple molecular architectures of the eye lens chaperone αB-crystallin elucidated by a triple hybrid approach. *Proc. Natl Acad. Sci. USA* **108,** 20491–20496 (2011).
16. Cenedella, R. J. *et al.* Direct perturbation of lens membrane structure may contribute to cataracts caused by U18666A, an oxidosqualene cyclase inhibitor. *J. Lipid Res.* **45,** 1232–1241 (2004).

Acknowledgements We thank the study participants for their support. We thank G. Hannum, M. Kruppa, J. H. Shin, J. Mei, H. Zheng, F. Xu, J. Zhang, M. Kircher, J. Shendure, S. J. Fliesler, J. Gleeson, X.-T. Zuo, Y. Li and Y. Ding for their helpful advice during the course of the experiments and data analysis. This work is supported in part by grants from 973 Project (2015CB94600, 2012CB917304), 863 Program (2014AA021604), NSFC (31327901), State Key Laboratory of Ophthalmology, and State Key Laboratory of Membrane Biology.

Author Contributions L.Zhao, Y.Liu., Y.-B.Y., L.Zhang and K.Z. designed the study, interpreted data and wrote the manuscript. L.Z., X.-J.C., J.Z., Y.-B.X., X.Y., L.-D.H, H.O., S.H.P., X.J., D.L., F.W., K.F., H.C., G.L., G.C., Y.Li, D.C., C.W., C.C., Y.W., A.Q., E.Y., W.W., X.H., S.G., Z.S., H.C.T., X.-J.Z., H.L., R.H., J.J.P.P., W.G., I.K., D.G., and X.S. performed the experiments; R.A., Y.Li and J.W. contributed to data analysis and interpretation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.Z. (kang.zhang@gmail.com), Y.-B.Y. (ybyan@tsinghua.edu.cn), Y.Z.L. (yzliu62@yahoo.com) or L.Zhang (zhang@ucsd.edu).

METHODS

Study participants. This study was approved by the Institutional Review Boards of Zhongshan Ophthalmic Center of Sun Yat-sen University, Sichuan University and the University of California, San Diego. Informed consent was obtained from all subjects before participation in the study. All participants underwent standard complete ophthalmic examinations and imaging studies. Demographic data, risk factors, and blood samples were collected at the initial visit. We recruited a consanguineous family consisting of two adults and four children. The parents were first cousins, and three of their four children were diagnosed with cataracts (Fig. 1a). We screened for *LSS* mutations in an additional 154 congenital cataract pedigrees and identified another family with a homozygous W581R mutation.

Exome capture and sequencing. Exome capture of phase I data (mother and three affected children) and phase II data (father and one unaffected daughter) were hybridized with NimbleGen 2.1M probe and Agilent SureSelect Human All Exon V2 according to the manufacturer's protocols, respectively. In brief, genomic DNA samples were randomly fragmented by Covaris with a base-pair peak of ~150–200 bp for the resulting fragments, and adapters were then ligated to both ends of the fragments. The adaptor-ligated templates were purified using Agencourt AMPure SPRI beads, and fragments with insert size ~250 bp were excised. Extracted DNA was amplified by ligation-mediated PCR, purified and hybridized to the SureSelect Biotinylated RNA Library (BAITS) for enrichment. Hybridized fragments bound to the strepavidin beads, whereas non-hybridized fragments were washed out after 24 h. Captured ligation-mediated PCR products were subjected to an Agilent 2100 Bioanalyzer to estimate the magnitude of enrichment. Each captured library was then loaded on Illunima Genome Analyzer II platform (phase I) or Hiseq 2000 (phase II), and paired-end sequencing was performed with read lengths of 90 bp, which provided at least 50× average coverage depth for each sample. Raw image files were processed by Illumina base-calling software for base calling with default parameters.

Read mapping and variant detection. Sequence reads in each individual were aligned to the human reference genome (NCBI build 37, hg19) using BWA[17](version 0.5.9–r16). BAM files created by BWA were then processed using the GATK[18] best practice pipeline using Genome Analysis ToolKit (version GATK 2.8) for re-alignment and variation (SNV and indel) detection. Variations that passed VQSR filtering criteria were extracted for the subsequent analyses.

Candidate causal variant identification. Variants were functionally annotated using ANNOVAR[19]. Missense, nonsense, and splicing mutations, which were likely to be deleterious when compared with synonymous and noncoding mutations, were extracted for analysis. Variants with a homozygous genotype in affected individuals and a heterozygous reference genotype in unaffected individuals, and a minor allele frequency $<0.5\%$ in both dbSNP137 and 1000 Genome Project databases (CEU) were considered as putative causal variants. Then, SIFT, Polyphen2, Phylop and Mutationtaster were used for function prediction, and mutations predicted to be damaging by no less than two tools were selected. Finally, variants that were shared as homozygous mutations within three affected children, as a heterozygous mutation in unaffected parents, and as heterozygous or homozygous to the reference genotype in unaffected daughter, but were absent in the public databases, were then considered as candidate causal variants.

Whole-genome genotyping. Whole-genome SNP genotyping was performed using Illumina HumanOmni 5Exome-4v1-1 array for all six family members. Mendelian error rate was calculated to check relative relationship in the family as part of quality control. Then, high-quality SNPs were selected for homozygosity mapping by HomozygosityMapper (http://www.homozygositymapper.org). HomozygosityMapper calculates the length of the homozygous block (in SNPs) at each marker for each sample. The values of the 'cases' are then added to get the 'homozygosity score' for a marker.

Mutation screening of *LSS* and *FDFT1* genes. Sanger DNA sequencing was performed to validate the G588S mutation in *LSS*. The 22 exons of the *LSS* gene were amplified by PCR and sequenced on the Genetic Analyzer 3130 (Applied Biosystems). The primers used to amplify the exons in *LSS* are presented in Extended Data Table 3a. We screened for mutations in the *LSS* gene in 154 families with congenital cataracts and identified another homozygous mutation, W581R, in a second consanguineous family. These two mutations were absent in 11,000 control chromosomes, including 2,000 chromosomes from an unaffected control population at the University of California, San Diego and the 1000 Genomes Project, and 8,000 chromosomes from an exome sequencing database at the University of Washington.

Due to a previous report that a *FDFT1* mutation modifies cataract phenotypes, we screened variants in the *FDFT1* gene, identifying only one common non-synonymous variant rs4731 (GRCh37/hg19: chr8:11666337; NM_001287742.1: c.134A > G, NM_001274671.1:p.K45R). The variant was excluded as the causal mutation since an unaffected daughter harboured the same homozygous change, and a relatively high frequency of general population possess this variant (minor allele frequency >4% in 1000 Genome Project data) (Extended Data Table 1e).

3D modelling of the G588S mutation. The model of the G588S mutant was built from two structures as determined by Ruf *et al.*[20] and deposited in the Protein Data Bank as entries 1W6K and 1W6J[12]. The X-ray coordinates were used to build a full-atom model of the enzyme, and it was refined using the Internal Coordinate Mechanics program (ICM) and its PDB conversion protocol[21]. To analyse the effect of the G588S-mutation-induced clash on lanosterol binding, we analysed all side chains involved in the pocket of the enzyme interacting with lanosterol using the 1W6K structure. The areas of contact were calculated as the differences between the solvent-accessible area of each residue with and without lanosterol and were sorted by size using the ICM program[22].

Plasmid constructs and site-directed mutagenesis. The clone containing *LSS* cDNA was purchased from Thermo Scientific Inc. The coding sequence of wild-type LSS was cloned and inserted into the pcDNA3.1-N-Flag plasmid (Invitrogen). The mutants were constructed via site-directed mutagenesis by overlap extension using PCR. The common PCR primers were: NdeI forward, 5′-CATATGACGGAGGGCACGTGTCT-3′ and XhoI reverse, 5′-CTCGAGTCAGGGGTGGCCAGCAAG-3′. The primers for constructing the W581R and G588S mutants were: W581R forward, 5′-TGGGAAGGCTCCCGGGGAGTTTGCT-3′; reverse, 5′-GTGAAGCAAACTCCCCGGGAGCCTTC-3′; G588S forward, 5′-GCTTCACCTACAGCACCTGGTTTG-3′; G588S reverse, 5′-CCAAACCAGGTGCTGTAGGTGAAG-3′. The recombinant pcDNA3.1-N-Flag plasmids containing the wild-type or mutated *LSS* genes were transformed into *E. coli* DH5α cells. The cDNA of αA-, αB-, βB2-, γC- and γD-crystallin were cloned from the total cDNA of human lens as described previously[23–26]. The mutants were constructed by site-directed mutagenesis using the primers listed in Extended Data Table 3b. The amplified fragments were digested by XhoI and BamHI, and then inserted into the eukaryotic expression vector peGFP-N1 or the prokaryotic expression vector pET28a. The plasmids were obtained using the Plasmid Maxiprep kit (Vigorous) and verified by DNA sequencing. Crystallin gene constructs were made as a C-terminus eGFP fusion protein, while LSS was made as an N-terminal Flag-tagged protein.

Cell culture and transfection. HeLa cells and human lens epithelial B-3 cells (HLEB-3) were obtained from ATCC. Human lens progenitor cells were isolated from a fetal human eye[27]. The HeLa cells were cultured in DMEM medium containing 10% FBS (Gibco). The HLEB-3 cells were cultured in F12 medium with 20% FBS, while human lens progenitor cells were cultured in MEM medium containing 20% FBS and 10 μg ml^{-1} FGF (Gibco). All cells were cultured at 37 °C in 5% CO_2 incubator. Cells routinely tested negative for mycoplasma contamination.

To assess the effect of *LSS* expression on sterol content, HeLa cells were transfected with wild-type *LSS* or *LSS* mutants fused with a Flag tag at the N-terminus of the coding region. The cells were harvested after 24 h transfection and the lipid fraction was extracted for LC–MS analysis. Cells transfected with the vector pcDNA3.1-N-Flag plasmids were used as a control. The expression levels of the wild-type and mutant LSS were normalized by western blot analysis using mouse anti-Flag (F1804; Sigma-Aldrich) and mouse anti-actin antibodies (BS6007M; Bioworld Technology).

To assess the effect of lanosterol on crystallin aggregation, human lens progenitor cells were co-transfected with *LSS* and various crystallin constructs for 4 h. Cells co-transfected with crystallin mutants and pcDNA3.1-N-Flag were used as a control. Human lens progenitor cells co-transfected with *LSS* and crystallin mutant constructs were cultured for 12 h before assaying for aggregates. The rescue experiments were performed after 16 h by addition of 40 μM sterols (lanosterol or cholesterol, Sigma-Aldrich) to the cell culture medium for 2 h, which was then replaced with fresh culture medium and cells cultured for 24 h. The percentage of cells with crystallin aggregates was calculated from ten randomly selected viewing fields. The values of the wild-type LSS group, mutant group, and mutant plus lanosterol group were calculated. Cells treated with 1% DMSO were used as the controls.

The impact of LSS and lanosterol on intracellular crystallin aggregation were evaluated in single-blinded observer studies. Experiments have been repeated at least three times. *P* values were calculated using Student's *t*-tests.

Fluorescence microscopy. Equal amounts of the human lens progenitor cells, HLEB-3 cells or HeLa cells were seeded on glass coverslips pretreated with TC (Solarbio). After culturing for 24 h to reach 90% confluency, the cells were transfected with plasmids containing various *LSS* or crystallin genes or co-transfected with plasmids containing a certain crystallin gene and those containing the wild-type or mutated *LSS* gene. The controls were cells transfected with the plasmids containing peGFP-N1 and/or peDNA3.1-N-Flag. Both transfection and co-transfection were performed using Lipofectamine 3000 (Invitrogen) according to the instructions from the manufacturer.

The effect of wild-type or mutated LSS on the intracellular aggregation of various cataract-causing crystallin mutants was evaluated by co-expression of

Flag–LSS and crystalline–GFP in the human lens progenitor cells, HLEB-3 cells or HeLa cells. The intracellular distributions of the proteins were visualized using GFP or antibody against Flag. After co-transfection for 4 h, the cells were cultured in fresh media for 24 h, and then analysed by microscopy.

The effect of lanosterol or cholesterol on the aggresome formation of various crystallins was studied by transfecting the cells with plasmids containing various crystallin genes. The cells were incubated for 24 h to enable efficient protein expression and aggresome formation. The cells were then treated with 0–40 μM sterols in 1% (for human lens progenitor cells) or 2% DMSO (for HeLa cells). Cells treated with 1% or 2% DMSO were used as the control. After treatment for 2 h, the media was replaced with fresh media. After 12 h, the cells were used for microscopy analysis.

The microscopy samples were prepared by washing the slips with phosphate buffered saline (PBS) three times. The cells were fixed with 4% paraformaldehyde for 40 min followed by another three washes with PBS. The cells were permeabilized with 0.1% Triton X-100 (Sigma) in PBS for 10 min and blocked with 5% normal goat serum in PBS for 1 h at 37 °C. Immunostaining was carried out by adding mouse anti-Flag antibody (1:500) or mouse anti-p62 antibody (1:200, ab56416; Abcam) in PBS buffer containing 5% normal goat serum and incubated for 1 h at 37 °C. Then the slips were washed three times with PBS, and further incubated with Alexa 649-conjugated goat anti-mouse IgG (1:250) for 1 h at ambient temperature. The nuclei were counterstained with Hoechst 33342 (Invitrogen). The mounted cells were analysed using a Carl Zeiss LSM 710 confocal microscope.

Live-cell imaging. Human lens progenitor cells were transfected with plasmids containing αA-crystallin(Y118D) mutant. After a 24 h transfection period, the cells with stable expression of αA-crystallin(Y118D) mutant were screen by incubation in culture medium containing 0.8 μg ml^{-1} G418 for 7 days. Then the obtained cells were seeded onto glass bottom cell culture dishes (In Vitro Scientific) and treated with 1% DMSO, 40 μM cholesterol in 1% DMSO or 40 μM lanosterol in 1% DMSO for 4 h. Fresh culture medium was added, and the cells were analysed by serial live-cell imaging. Live-cell images were viewed with an Olympus IX81 microscope and captured with CellSens Dimension software (Olympus). Quantitative analysis of the size of aggregates was performed by measuring the fluorescence intensity of p62-positive aggregates using single-particle tracking in live-cell imaging. The live-cell imaging was conducted using three biological replicates with 1–8 repetitions each.

Lipid extraction of the cells. Extraction of lipids was performed using the Bligh and Dyer method[28]. In brief, $\sim 1 \times 10^6$–10^7 HeLa cells were washed 3–5 times with PBS and then scraped in 400-μl ice-cold methanol and transferred to a 1.5 ml Eppendorf tube with the addition of 200 μl chloroform. The samples were vortex-agitated for 1 min and then mixed with 300 μl of 1 M KCl. The organic and aqueous phases were separated by microcentrifugation at 20,817*g*. for 5 min at 4 °C. After separation, the lower organic phase was collected. Then the residual aqueous phase was re-extracted twice using 300 μl chloroform. The collected organic phases were dried using a SpeedVac sample concentrator under vacuum. The dried samples were stored at −80 °C for further LC–MS analysis.

LC–MS analysis. The dried lipid extracts were re-suspended in 100 μl methanol. The samples were vortex-agitated for 10 min, treated by 80 W ultrasonic sonication for 30 min, microcentrifuged at 20,817*g* for 10 min, and then the supernatant was transferred to a new Eppendorf tube. The microcentrifugation treatment was repeated three times. The derived samples were analysed by an Agilent 1290/6460 triple quadrupole LC/MS using an alternative Atmospheric Pressure Chemical Ionisation (APCI) source. The lipids were separated using an Agilent SB-C18 column. Selective ion monitoring was performed using the electron ionization mode. The highly pure lanosterol and cholesterol were used as controls. The MS determination was performed using a gas temperature of 350 °C, a gas flow rate of 4 l min^{-1}, a nebulizer of 60 p.s.i., a vaporizer of 350 °C, a capillary of 3,500 V and a corona current of 4 μA. To optimize the sensitivity and specificity, two qualifier ions were selected for the MS analysis of each compound (369.3/161.1 and 369.3/147 for cholesterol, and 409.2/191.3 and 409.2/109 for lanosterol).

Western blotting. The cell lysates were prepared in RIPA buffer containing 50 mM Tris (pH 8.0), 150 mM NaCl, 1% Triton X-100, 1 mM EDTA, 0.5% sodium deoxycholate and 0.1% SDS. The supernatant and precipitation fractions were separated by centrifugation. The proteins were separated by a 12.5% SDS–PAGE and transferred to a PVDF membrane (GE Healthcare). The mouse antibodies against Flag (F1804; Sigma-Aldrich) or GFP (MB2005; Bioworld Technology) were used to identify the overexpressed LSS and crystallin proteins, respectively. Quantification of the western blot bands was achieved using the software GELPRO (Media Cybernetics). The presented quantitative data were calculated from three independent experiments.

Protein expression and purification. The recombinant His-tagged wild-type and mutated β- and γ-crystallin proteins were overexpressed in *Escherichia coli* Rosetta and purified using an Ni-NTA affinity column followed by gel filtration chromatography using the same protocol as described elsewhere[23,24,26,29]. The overexpression and purification of the non-tagged αA- and αB-crystallins were performed as described previously[30]. The purity of the proteins was estimated to be above 95% as evaluated by one homogeneous band on 12.5% SDS–PAGE, 10% native-PAGE and a single peak in the size-exclusion chromatography profile. The protein concentration was determined according to the Bradford method by using BSA as the standard[31]. All protein samples were prepared in 20 mM PBS buffer containing 150 mM NaCl, 1 mM EDTA and 1 mM DTT.

Protein aggregation and aggregate dissociation. The aggregates of the wild-type and mutated αA- and αB-crystallin proteins were obtained by heating the protein solutions containing 1 M guanidine chloride (ultrapure, Sigma-Aldrich) at a concentration of 5 mg ml^{-1} at 60 °C for 2 h. The aggregates of the wild-type and mutated β- and γ-crystallins were prepared by heating the protein solutions containing 1 M guanidine chloride at 37 °C for 48 h. The formation of aggregates was confirmed by ThT fluorescence, turbidity (absorbance at 400 nm) and transmission electron microscopy (TEM) observations. The preformed aggregates were re-suspended in 20 mM PBS with a final concentration of 0.2 mg ml^{-1} (approximately 10 μM). The re-suspended aggregates were treated with 500 μM lanosterol or cholesterol in liposomes formed by 500 μM DPPC (Sigma-Aldrich) at 37 °C. Aggregates treated by 500 μM DPPC liposome were used as a negative control. After 24 h of treatment, the protein solutions were used for ThT fluorescence, turbidity and negatively stained TEM observations. The TEM samples were prepared by depositing the protein solutions onto a freshly glow-discharged carbon-coated copper grid. Negative-staining samples were obtained by staining the grid with 1.25% uranyl acetate for 30 s. The negatively stained TEM pictures were obtained on a Hitachi H-7650B transmission electron microscope with a voltage of 120 kV and a magnification of 48,000.

Treatment of cataractous rabbit lenses. This study was approved by IACUC of Zhongshan Ophthalmic Center and West China Hospital. Rabbits were euthanized by CO_2 inhalation and lenses were immediately dissected and treated with vehicle or lanosterol dissolved in vehicle to make 25 mM solutions. Lens tissues were incubated in these solutions for 6 days in the dark at room temperature. Cataracts were examined under a microscope and photographed. Degree of cataract was assessed by a blinded examiner using a previously described opacification grading system, shown below[32,33]. Improvements in lens clarity and transparency were quantified by visual inspection and grading. Lens clarity was scored by transmission of light, clarity of a grid image underneath the lense (Extended Data Fig. 7), and improvement in overall clarity of a lens or improvement in clarity of localized areas of cortical cataract. Wilcoxon test was used to evaluate the treatment effect.

Cataract grading system.

- Grade 0: absence of opacification (gridlines clearly visible);
- Grade 1: a slight degree of opacification (minimal clouding of gridlines, with gridlines still visible);
- Grade 2: presence of diffuse opacification involving almost the entire lens (moderate clouding of gridlines, with main gridlines visible);
- Grade 3: presence of extensive, thick opacification involving the entire lens (total clouding of gridlines, with gridlines not seen at all)

Preparation of drug-loaded nanoparticles. Lanosterol was loaded into a lipid-polymer hybrid nanoparticle through an adapted nanoprecipitation method[34]. In brief, the desired concentration of lanosterol was mixed with polycaprolactone (PCL) polymer dissolved in acetonitrile. Lecithin and 1,2-distearoyl-*sn*-glycero-3-phosphoethanolamine-*N*-carboxy(polyethylene glycol) 2000 (DSPE-PEG-COOH) were dissolved in a 4% ethanol aqueous solution at 20% of the PCL polymer weight and heated above 60 °C. The lanosterol/PCL solution was then added into the preheated lipid solution under gentle stirring followed by rigorous vortexing for 3 min. The mixture solution was then stirred for 2 h to allow the nanoparticles to form and the acetonitrile to evaporate. Next, the nanoparticle solution was washed three times using an Amicon Ultra-4 centrifugal filter (Millipore) with a molecular weight cut-off of 10 kDa to remove the remaining organic solvent and free molecules. The resulting nanoparticles were then re-suspended in PBS buffer for subsequent use. The size, size distribution, and surface zeta potential of the drug-loaded nanoparticles were characterized by dynamic light scattering. The loading yield of lanosterol was quantified by high-performance liquid chromatography.

Treatment of cataractous lenses in dogs. This study was approved by IACUC of Zhongshan Ophthalmic Center and West China Hospital. The following adult dog breeds were used for assessing the treatment effect: black Labrador, Queensland Heeler, Miniature Pinscher. All dogs were adult, non-diabetic and had normal ocular surfaces and ocular adnexa, with naturally occurring adult onset cataracts. There were near equal distributions of male and female dogs. We screened all exons of the *LSS* gene in these dogs and did not find any mutations. To assess the effect of lanosterol treatment on cataracts in live animals, dogs were pre-medicated with intramuscular injections of acepromaxine and butorphanol. After 20 min, induction of anaesthesia was performed by application of intravenous propofol. Dogs were then immediately intubated and maintained on oxygen and 2% isoflurane at 2 l min^{-1}. Lanosterol (100 μg)-loaded nanoparticles were initially injected into the vitreous cavity in the test eye using a 28-gauge needle, and then were given every 3 days for the duration of the experiment. Treatment eyes or sham eyes were randomized. The control eye was given an injection with empty nanoparticle carriers as a negative control. The treatment eyes were treated with lanosterol in topical eye drops (see below for eye drop formulation). One 50-μl drop of lanosterol was administered three times daily to the test eye over 6 weeks. Degree of cataract severity was examined by slit lamp and photographed at the beginning and the end of the 6-week treatment period. Prior to examinations, pupils were dilated with 1% tropicamide and 10% phenylephrine. Degree of cataract severity was assessed by a blinded examiner and scored based on canine cataract stage, shown below[35]. Improvements in lens clarity and transparency were quantified. Wilcoxon test was used to evaluate the treatment effect.

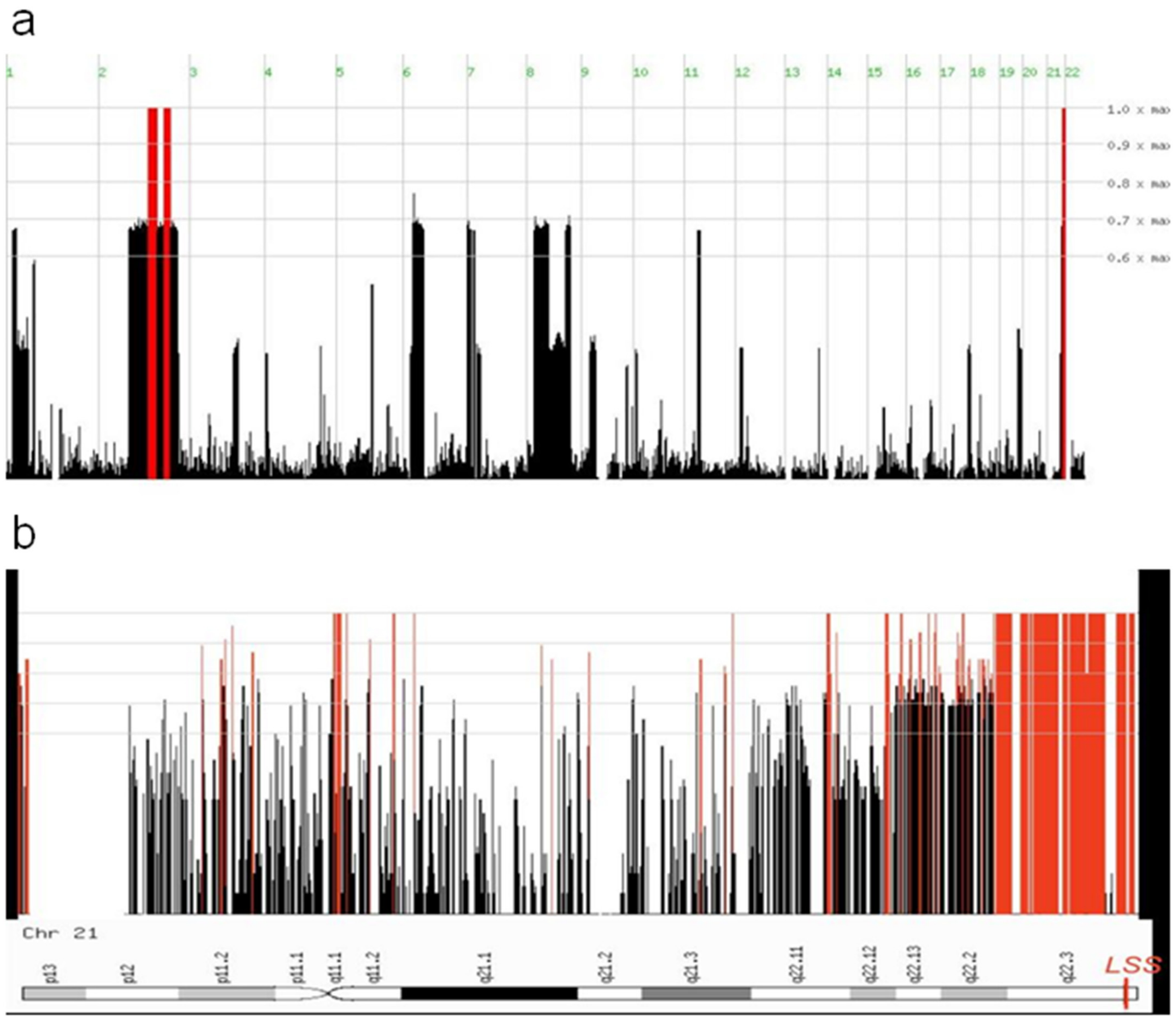
Grading system of canine cataracts.

- Grade 0: absence of opacification (no cataract);
- Grade 1: a slight degree of opacification (incipient stage);
- Grade 2: presence of diffuse opacification involving almost the entire lens (immature stage);
- Grade 3: presence of extensive, thick opacification involving the entire lens (mature stage)

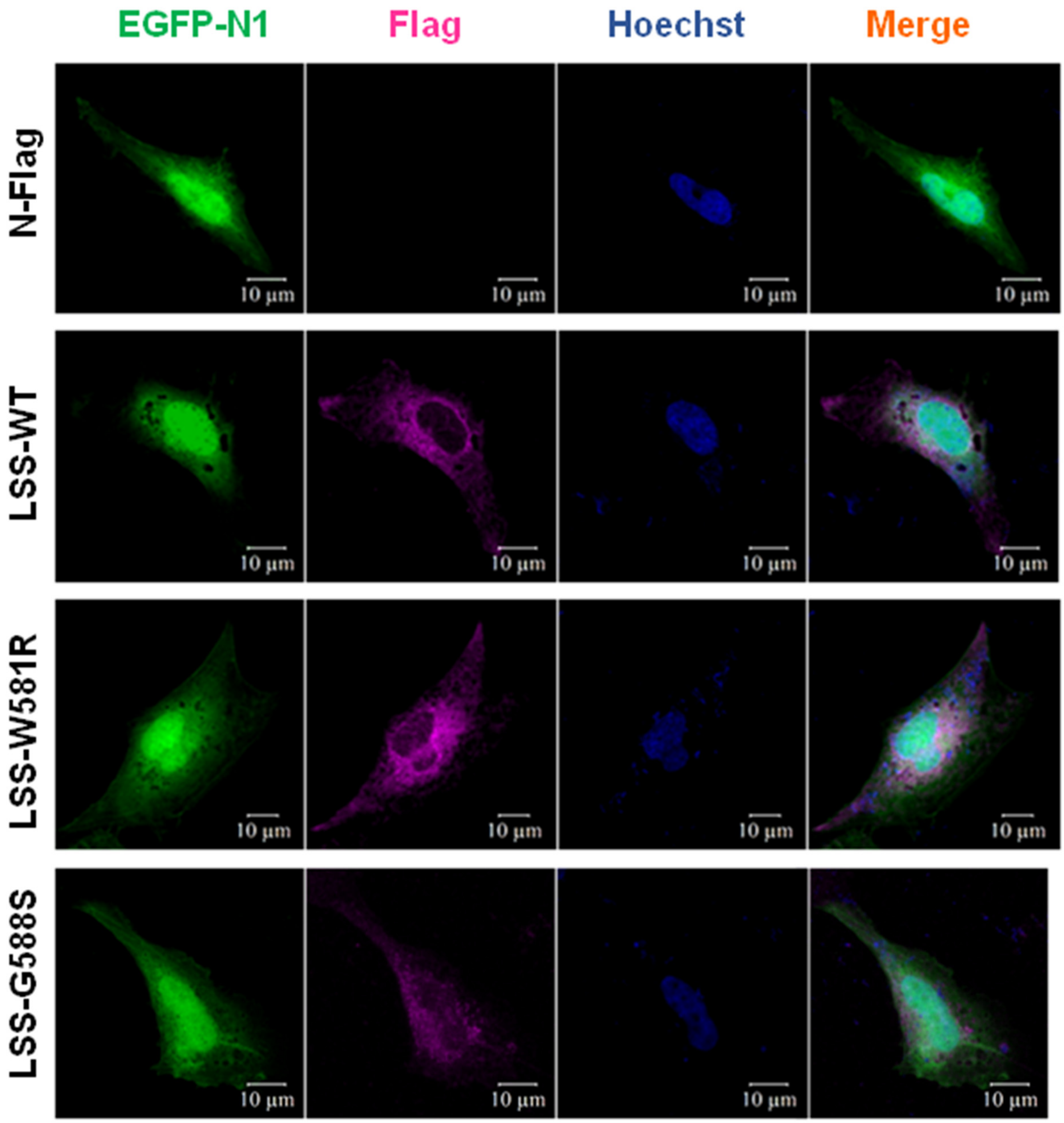
Topical vehicle solution. Double distilled H_2O was added to 1.1 g $(EDTA)_2Na$ combined with 0.055 g alkyldimethylbenzylammonium chloride until a final volume of 1.1 l (pH 5.66) was achieved.

25 mM lanosterol in the topical vehicle solution. Double distilled H_2O was added to a mixture of 12.5 g lanosterol, 1.1 g $(EDTA)_2Na$, 0.055 g alkyldimethylbenzylammonium chloride and 200 ml EtOH to a final volume of 1.1 l.

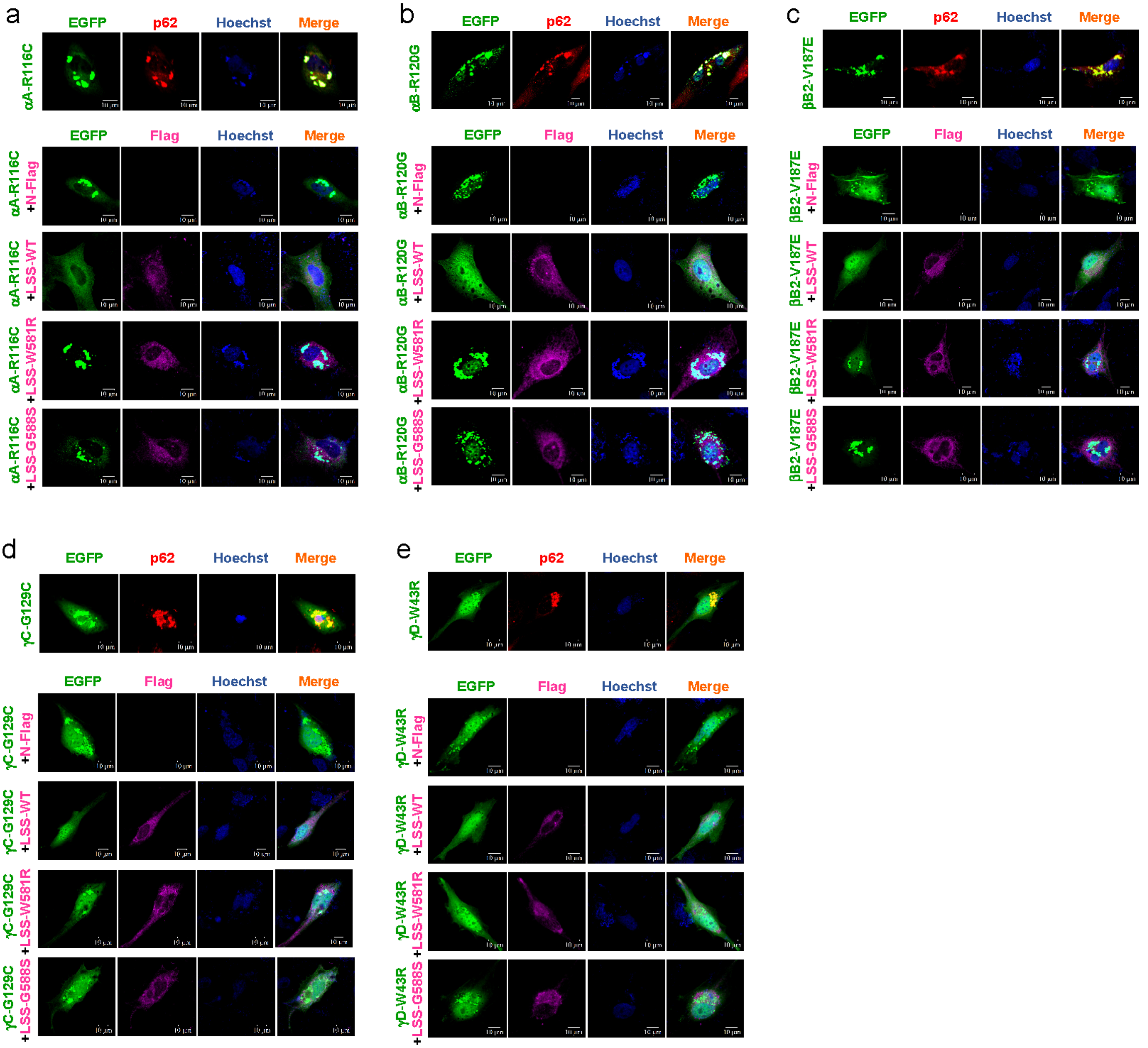
17. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
18. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43,** 491–498 (2011).
19. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).
20. Ruf, A. *et al.* The monotopic membrane protein human oxidosqualene cyclase is active as monomer. *Biochem. Biophys. Res. Commun.* **315,** 247–254 (2004).
21. Cardozo, T., Totrov, M. & Abagyan, R. Homology modeling by the ICM method. *Proteins* **23,** 403–414 (1995).
22. Abagyan, R. & Argos, P. Optimal protocol and trajectory visualization for conformational searches of peptides and proteins. *J. Mol. Biol.* **225,** 519–532 (1992).
23. Xu, J. *et al.* The congenital cataract-linked A2V mutation impairs tetramer formation and promotes aggregation of βB2-crystallin. *PLoS ONE* **7,** e51200 (2012).
24. Wang, B. *et al.* A novel *CRYGD* mutation (p.Trp43Arg) causing autosomal dominant congenital cataract in a Chinese family. *Hum. Mutat.* **32,** E1939–E1947 (2011).
25. Gu, F. *et al.* A novel mutation in *AlphaA-crystallin (CRYAA)* caused autosomal dominant congenital cataract in a large Chinese family. *Hum. Mutat.* **29,** 769 (2008).
26. Li, X.-Q. *et al.* A novel mutation impairing the tertiary structure and stability of γC-crystallin (CRYGC) leads to cataract formation in humans and zebrafish lens. *Hum. Mutat.* **33,** 391–401 (2012).
27. Nagineni, C. N. & Bhat, S. P. Human fetal lens epithelial cells in culture: an in vitro model for the study of crystallin expression and lens differentiation. *Curr. Eye Res.* **8,** 285–291 (1989).
28. Bligh, E. G. & Dyer, W. J. A rapid method of total lipid extraction and purification. *Can. J. Biochem. Physiol.* **37,** 911–917 (1959).
29. Wang, S., Leng, X.-Y. & Yan, Y.-B. The benefits of being β-crystallin heteromers: βB1-crystallin protects βA3-crystallin against aggregation during co-refolding. *Biochemistry* **50,** 10451–10461 (2011).
30. Sun, T.-X., Das, B. K. & Liang, J. J. N. Conformational and functional differences between recombinant human lens αA- and αB-crystallin. *J. Biol. Chem.* **272,** 6220–6225 (1997).
31. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72,** 248–254 (1976).
32. Geraldine, P. *et al.* Prevention of selenite-induced cataractogenesis by acetyl-L-carnitine: an experimental study. *Exp. Eye Res.* **83,** 1340–1349 (2006).
33. Makri, O. E., Ferlemi, A. V., Lamari, F. N. & Georgakopoulos, C. D. Saffron administration prevents selenite-induced cataractogenesis. *Mol. Vis.* **19,** 1188–1197 (2013).
34. Zhang, L. *et al.* Self-assembled lipid–polymer hybrid nanoparticles: a robust drug delivery platform. *ACS Nano* **2,** 1696–1702 (2008).
35. La Croix, N. Cataracts: When to refer. *Top. Companion Anim. Med.* **23,** 46–50 (2008).



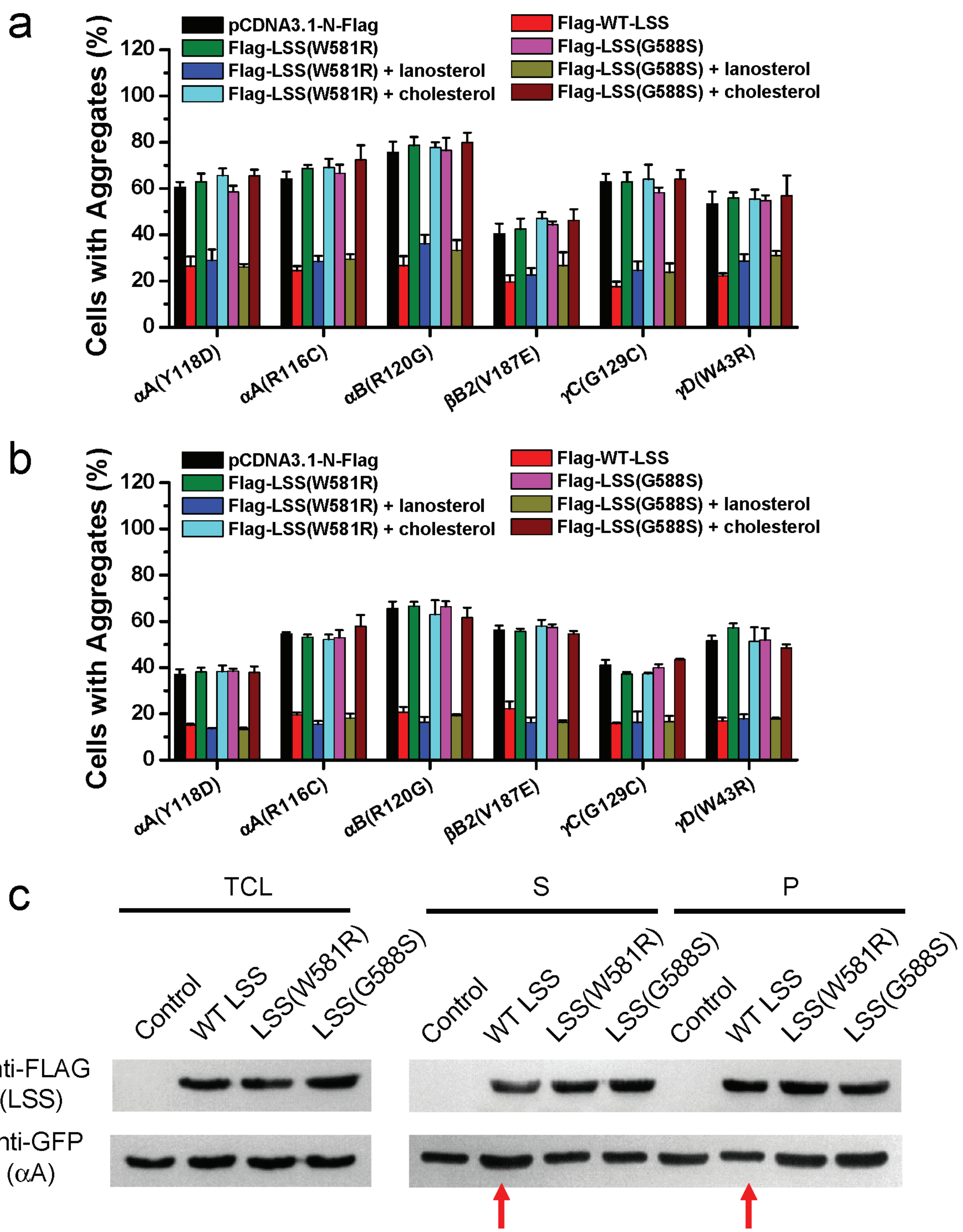
Extended Data Figure 1 | Genome-wide homozygosity. **a**, HomozygosityMapper plots the genome-wide homozygosity as bar charts. To emphasize regions of interest, any score higher than 80% of the maximum score reached in this project is coloured in red. **b**, The homozygosity scores were plotted against the physical position on chromosome 21, which contains the *LSS* gene. Red bars indicate regions with highest scores. The right side of the chromosome contains a long continuous homozygous region, where the *LSS* gene is located.



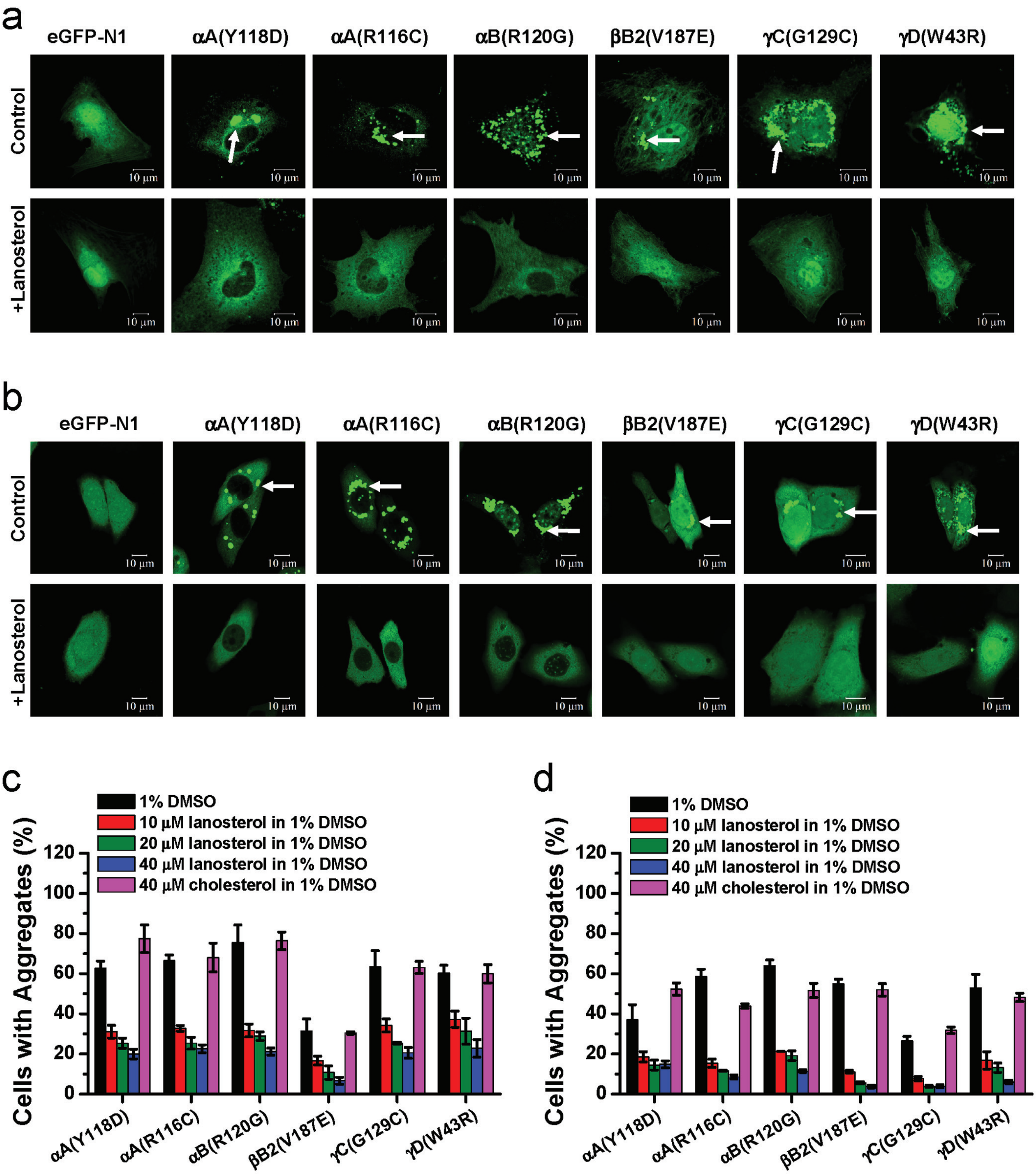
Extended Data Figure 2 | Representative confocal images of cells co-transfected with Flag–LSS and eGFP. Human lens progenitor cells were co-transfected with either the wild-type or the mutated *LSS* gene and the *eGFP* gene for 4 h and cultured for 16 h in fresh culture medium. The cellular distribution of LSS was then visualized using an anti-Flag antibody (purple). The distribution of eGFP (green) was used as a control. The nuclei were stained and visualized by Hoechst 33342 (blue).



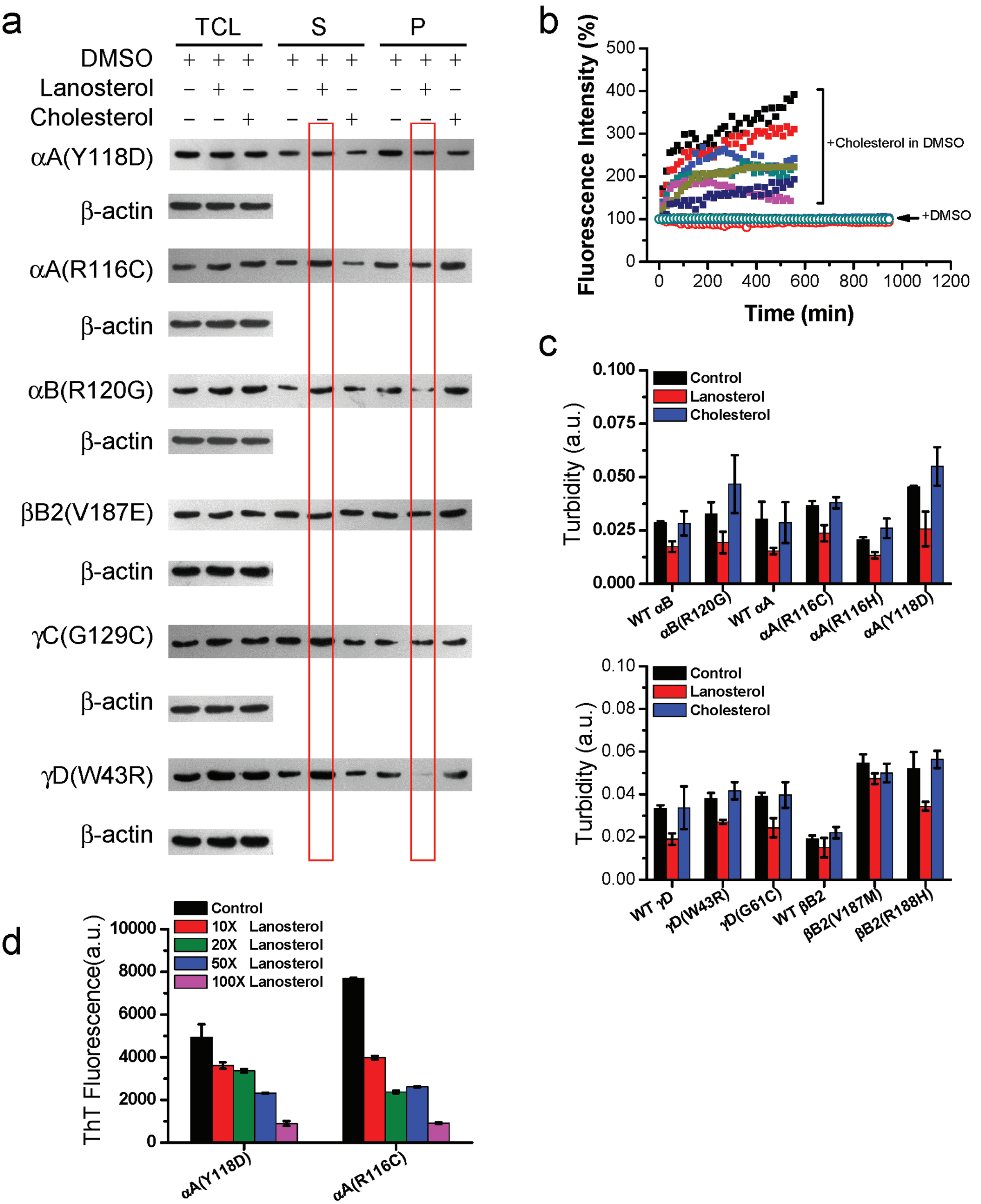
Extended Data Figure 3 | Representative confocal images of cells co-transfected with LSS and various cataract-causing crystallin mutants. **a**, R116C mutant of αA-crystallin. **b**, R120G mutant of αB-crystallin. **c**, V187E mutant of βB2-crystallin. **c**, G129C mutant of γC-crystallin. **e**, W43R mutant of γD-crystallin. Human lens progenitor cells were co-transfected with either the wild-type or the mutated Flag–LSS gene and the mutant GFP–crystallin gene for 4 h and cultured for 16 h in fresh culture medium. All crystallin mutants formed p62-positive aggregates as indicated by the co-localization of the mutant crystallins and p62. Cells co-transfected with GFP–crystallin and pcDNA3.1-N-Flag were used as controls. The formation of intracellular aggregates of various crystallin proteins was visualized by fluorescence of GFP (green). Wild-type or mutated LSS was detected with an anti-Flag antibody (purple), p62 was stained using an anti-p62 antibody (red), while the nuclei were stained and visualized by Hoechst 33342 staining (blue). Quantitative analysis of cells with aggregates is summarized in Fig. 3c.



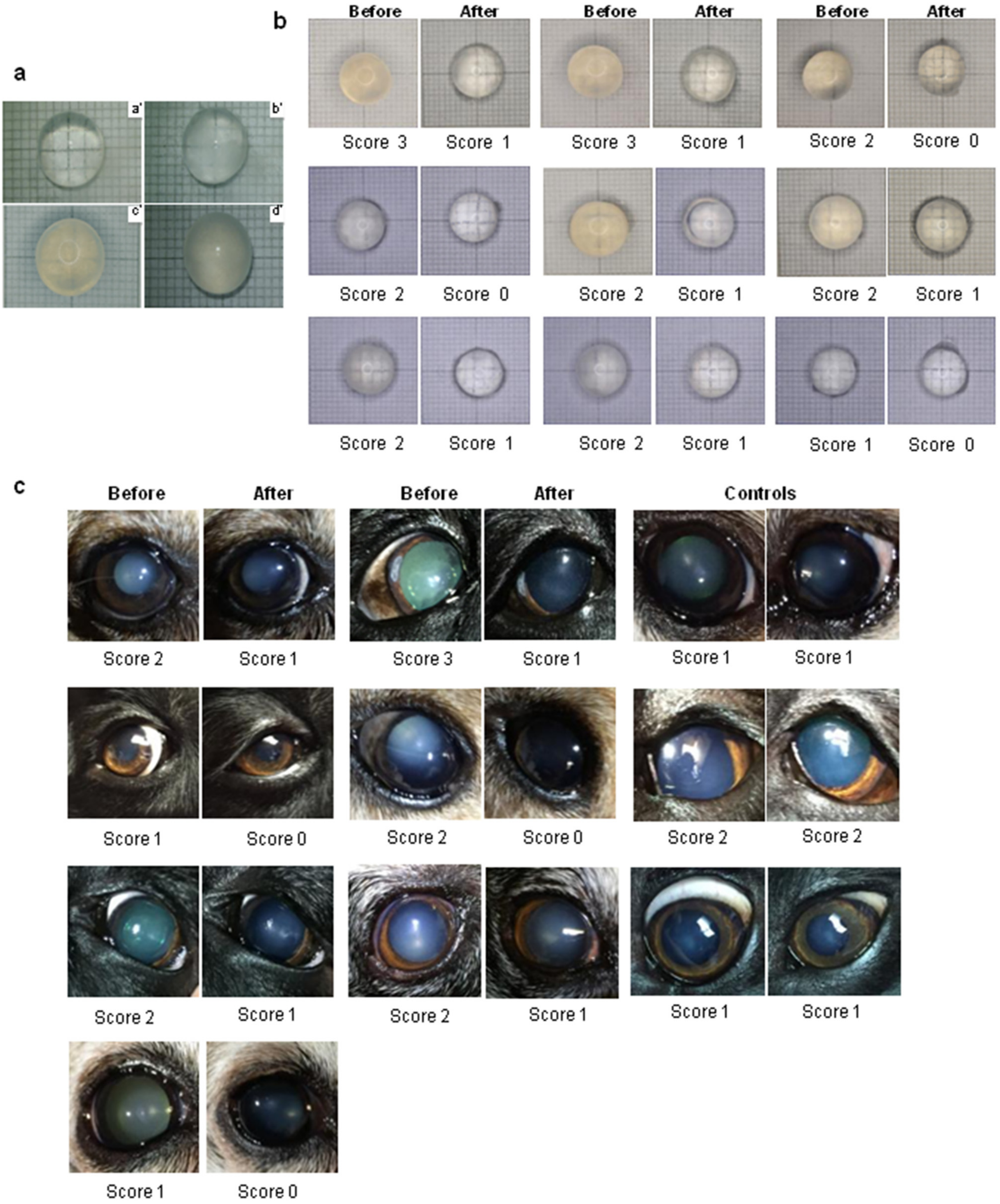
Extended Data Figure 4 | Inhibition of crystallin mutant aggregation by wild-type LSS and lanosterol in HLEB-3 cells (a) or HeLa cells (b). Cells co-transfected with LSS and crystallin mutant constructs were cultured for 24 h before assaying for aggregates. The rescue experiments were performed by addition of 40 μM sterols (lanosterol or cholesterol) to the cell culture medium for 2 h, the sterol medium was then replaced with fresh culture medium and the cells were cultured for a further 12 h. The percentage of cells with crystallin aggregates were calculated from ten randomly selected viewing fields. The values of the wild-type LSS group, mutant group, or mutant plus lanosterol group were calculated. Aggregates were significantly lower in the wild-type LSS and lanosterol groups compared to the control group ($P < 1 \times 10^{-4}$), while aggregates in mutant LSS or cholesterol groups showed no difference to the control group ($P > 0.1$). **c**, Human lens progenitor cells were co-transfected with wild-type or mutant LSS plus αA-crystallin(Y118D). αA-crystallin(Y118D) co-expressed with pcDNA3.1-N-Flag was used as a control. After transfection for 4 h and incubation in fresh culture medium for another 24 h, the cells were lysed and centrifuged to separate supernatant and insoluble fractions. LSS and crystallin fusion proteins were detected by antibodies against Flag and GFP, respectively. Red arrows indicate higher crystalline content in the soluble fraction versus in the insoluble fraction in cells containing the WT-LSS. Data were quantified from three independent experiments and summarized in Fig. 3d.



Extended Data Figure 5 | Lanosterol significantly reduced the intracellular aggregation caused by various cataract-causing mutant crystallin proteins in a concentration-dependent manner when assayed in HLEB-3 or HeLa cells. **a**, Representative confocal images of HLEB-3 cells transfected with various cataract-causing crystallin mutants. **b**, Representative confocal images of HeLa cells transfected with various cataract-causing crystallin mutants. Cells were transfected with various crystallin constructs for 4 h and cultured for an additional 24 h in fresh culture medium. Then the cells were treated with 10, 20 and 40 µM lanosterol in 1% (HLEB-3 cells) or 2% DMSO (HeLa cells) for 2 h and cultured for another 12 h. Cells treated with 1% (HLEB-3 cells) or 2% DMSO (HeLa cells) were used as the controls. Formation of intracellular aggregates of various crystallin proteins was visualized by fluorescence of GFP (green) and the nuclei were stained with Hoechst 33342 (blue). Typical intracellular aggregates are indicated by arrows. **c**, Concentration dependence of the aggregation-dissolving effects of lanosterol when assayed in HLEB-3 cells. **d**, Concentration dependence of the aggregation-dissolving effects of lanosterol when assayed in HeLa cells.



Extended Data Figure 6 | Treatment by lanosterol, but not cholesterol, increased cataract-causing mutant crystallins in soluble fractions when compared to a control group or a mutant LSS group. **a**, Human lens progenitor cells were transfected with mutant crystallin genes for 4 h, and then incubated in fresh culture medium for another 24 h. The cells were harvested and lysed. Supernatant and insoluble fractions were separated by centrifugation and analysed by western blot analysis. LSS and crystallin fusion proteins were identified by antibodies against Flag and GFP tags, respectively. The lanosterol-treated group is highlighted by red boxes. Cells treated with 1% DMSO were used as a control. β-Actin was used as an internal protein loading control of total cell lysates (TCL). S, supernatant; P, insoluble fraction. **b**, Effect of DMSO (*n* = 4) and cholesterol (*n* = 7) on the size changes of αA-crystallin(Y118D) aggregates in human lens progenitor cells evaluated by single-particle tracking in live-cell imaging. **c**, Evaluation of the effect of lanosterol on the dissolution of crystallin aggregates by turbidity. Crystallin aggregates were formed by incubating 5 mg ml^{-1} protein solution at 60 °C for 2 h (α-crystallins) or 37 °C for 48 h (β- and γ-crystallins) in the presence of 1 M guanidine chloride. The preformed aggregates were re-suspended in PBS at a final protein concentration of 0.2 mg ml^{-1} and were treated with 500 μM sterols in 500 μM DPPC liposome and incubated at 37 °C for 24 h. Aggregates treated with 500 μM DPPC liposome only were used as the controls. **d**, Concentration-dependent effect of lanosterol on the re-dissolution of amyloid-like fibrils by αA-crystallin mutants evaluated by ThT fluorescence. Aggregates treated with 500 μM DPPC liposome only were used as the controls.



Extended Data Figure 7 | Grading system of cataractous lenses. **a**, Lenses were placed above a grid and photographed. The degree of transparency was scored as 0, a clear lens and absence of opacification (gridlines clearly visible, a′); 1, a blurry lens and a slight degree of opacification (minimal clouding of gridlines, with gridlines still visible, b′); 2, a cloudy lens and presence of diffuse opacification involving almost the entire lens (moderate clouding of gridlines, with main gridlines visible, c′); or 3, an opaque lens and presence of extensive thick opacification involving the entire lens (total clouding of gridlines, with gridlines not seen at all, d′). **b**, Lanosterol reduced cataract severity and increased clarity in isolated cataractous rabbit lenses. Rabbit lenses ($n = 13$) were dissected and incubated with lanosterol for 6 days and subsequently assessed for lens clarity and transparency. Pairs of photographs of each cataractous rabbit lens showing before and after treatment with scores underneath are shown. **c**, Lanosterol reduced cataract severity and increased lens clarity in dogs. Dog eyes with cataracts ($n = 7$) were treated with lanosterol for 6 weeks and assessed for lens clarity and transparency. A pair of photographs of each study eye before and after treatment is shown with scores underneath. Three control eyes treated with vehicles alone are also presented.

Extended Data Table 1 | Exome sequencing and variants

a

Sample	Total effective yield(Mb)	Average sequencing depth	Mismatch rate	Coverage of target region	Fraction of target covered >= 4x	Fraction of target covered >= 10x
IV-1	3,409.20	60.16	0.20%	99.60%	99.10%	97.60%
IV-2	3,314.58	58.62	0.20%	99.60%	99.20%	97.80%
IV-3	3,327.63	57.24	0.20%	99.80%	99.20%	97.40%
III-2	3,029.40	51.89	0.21%	99.80%	99.30%	97.70%
III-1	6,877.08	54.24	0.29%	96.30%	89.40%	81.80%
IV-4	6,331.78	44.12	0.29%	96.50%	88.80%	79.80%

b

Sample	Total variation	Heterzygotes	Homozygotes	missense	nonsense	readthrough	synonymous	splicing	intergenic	intronic
IV-1	61,189	35,571	25,618	6,105	69	39	7,296	32	5,371	36,598
IV-2	60,829	34,698	26,131	6,074	62	41	7,211	38	5,178	36,572
IV-3	61,078	35,238	25,840	6,221	78	43	7,265	38	5,099	36,544
III-2	62,753	39,001	23,752	6,393	64	38	7,588	34	5,764	36,924
III-1	80,067	49,694	30,373	7,247	93	49	8,166	47	15,063	41,391
IV-4	80,893	48,211	32,682	7,252	85	50	8,184	50	14,547	42,414

c

Filters	III-1 (carrier father)	III-2 (carrier mother)	IV-1 (affected daughter)	IV-2 (affected son)	IV-3 (affected son)	IV-4 (unaffected daughter)	Combine
Total variations	80,067	62,753	61,189	60,829	61,078	80,893	-
Missense, Nonsense, Splicing	7,389	6,495	6,213	6,177	6,342	7,387	-
Affected: 1/1; carrier: 0/1; unaffected: 0/1 or 0/0 *	5,792	4,661	3,127	3,123	3,085	5,638	9
Not in dbSNP	3,724	2,969	1,954	1,929	1,928	3,589	5
Not in 1000 Genomes Project	1,032	767	227	264	245	1,059	1
Predicted damaging	267	269	31	45	41	264	1

*Homozygous in affected child, heterozygous in carrier, no homozygous mutants in unaffected child

d

Sample	Total loci	Captured	SNP
IV-1	4,641,218	4,440,318	559,832
IV-2	4,641,218	4,446,992	605,499
IV-3	4,641,218	4,445,267	526,794
III-2	4,641,218	4,448,054	537,925
III-1	4,641,218	4,446,581	574,880
IV-4	4,641,218	4,450,657	584,347

e

Position (GRch37/hg19)	refSNP	REF	ALT	Function	III-1 (carrier father)	III-2 (carrier mother)	IV-1 (affected daughter)	IV-2 (affected son)	IV-3 (affected son)	IV-4 (unaffected daughter)
chr8:11666337	rs4731	A	G	nonsynonymous	A/G	A/G	G/G	A/G	G/G	G/G
chr8:11683653	rs904011	T	C	synonymous	C/C	C/C	C/C	C/C	C/C	C/C

a, Summary of exome sequencing data production. **b**, Summary of detected variants. **c**, Variants prioritization pipeline after exome sequencing. **d**, Summary of whole-genome genotyping data. **e**, Coding variants detected on gene *FDFT1*.

Extended Data Table 2 | Treatment effect of lanosterol in cataractous rabbit lenses and dog cataracts.

a

Sample number	Before treatment	After treatment
1	3	1
2	2	0
3	2	1
4	2	0
5	3	1
6	2	1
7	2	1
8	2	0
9	1	1
10	1	0
11	2	1
12	1	1
13	2	1

Grading of the cataract severity was conducted by an examiner, blinded regarding the treatment status, based on a scale 0 to 3, as described in the Methods section.

b

Study eye	Treatment group		Control group	
	Before	After	Before	After
1	2	1	1	1
2	1	0	2	2
3	2	1	1	1
4	3	1		
5	1	0		
6	2	0		
7	2	1		

Grading of the cataract formation was conducted by an examiner, blinded regarding the treatment status, based on a scale 0 to 3. 0 = no cataract, 1 = incipient, 2 = immature, and 3 = mature.

a, Treatment effect of lanosterol in cataractous rabbit lenses. Grading of the cataract severity was conducted by an examiner, blinded to treatment status, on a scale from 0 to 3, as described in the Methods. **b**, Treatment effect of lanosterol in cataractous dog lenses. Grading of the cataract formation was conducted by an examiner, blinded to treatment status, on a scale from 0 to 3. 0 = no cataract, 1 = incipient, 2 = immature, and 3 = mature.

Extended Data Table 3 | Primers used for sequencing of each exon in the human *LSS* gene and construction of crystallin mutants

a

Amplicon	Sequence (5'-3')
LSS-Exon1-F	GCCTGAGCGCCTGCCGAGGCCT
LSS-Exon1-R	GACACCTGAGGACCACCGGCCAT
LSS-Exon2-F	GTGGTCCTAGGTGCTGAGGAGA
LSS-Exon2-R	CGTGCTCCTCACGGCTCACCCCT
LSS-Exon3-F	CTTGGGCTGTATGTGAAGAGGGT
LSS-Exon3-R	CCTAGACCAGGCTGGGCCAGGAT
LSS-Exon4-F	GTTGGAGTGAGGTGCTCAGGAGGA
LSS-Exon4-R	GCAGCTGCCTGGAAACCCAAGCAT
LSS-Exon5-F	GCATTCTTAGTTTTCTGAGGAAACTC
LSS-Exon5-R	CCACTGTTTCAGCTGCAAGTGCAT
LSS-Exon6-F	CAGAGGGTGAAGCTTCCCAGCT
LSS-Exon6-R	GCTGTCACAGCCTGCACCTGAC
LSS-Exon7-F	GAAAGGGCCCAAGGTATGGATGCT
LSS-Exon7-R	GTGAGTGGACAGGTGTGGTTAGAT
LSS-Exon8-F	GAGCCAGGCCTACCAGGTGCT
LSS-Exon8-R	GCAGGGGATGAGTGCGTGAAT
LSS-Exon9-F	GCAGTGCATGGAGCTCCAGGCT
LSS-Exon9-R	CCAGGAAACCCCACTCCCAGCT
LSS-Exon10-F	GTGGATCTGGACGAGACCTTGT
LSS-Exon10-R	CACTGGGATGCAGCTGGGGCT
LSS-Exon11-F	GTGCAGGGTCTGGGTAGCAGCT
LSS-Exon11-R	GACATGATTGCAAAGGAAGCAT
LSS-Exon12-F	CTGGAGGCAGTGGCTGGGAGT
LSS-Exon12-R	GCAAGTGTGTGGCCAGCAGTGCT
LSS-Exon13-F	GGCAGGATGTGGCCAGGACCAT
LSS-Exon13-R	GCACTTCTGCCTGCAGGAGCT
LSS-Exon14-F	CCAGTCTGTCTCAGCGATGT
LSS-Exon14-R	CCAAAAACGCCAAGGGAGGAGT
LSS-Exon15-F	CTGGCTGCACCCACACCTTTGGT
LSS-Exon15-R	GCTCATCTGCAGGACACGAGGT
LSS-Exon16-F	GTTGTCAGCCCTAGTGTTGCCT
LSS-Exon16-R	CAGGTTTGTGTACCACAGTGCT
LSS-Exon17-F	GAGCTGCAGAGCCTGGGCAGCCA
LSS-Exon17-R	CCGTGTCACAGAATGATGCGT
LSS-Exon18-F	GAATTGGGATAGGTAAACTGCT
LSS-Exon18-R	CGCAGTGTGTGAGAGCAGAAACCT
LSS-Exon19-F	CTTAATGCCTGAGGCACTGGAGT
LSS-Exon19-R	CACTCATGACAGAGCATTGGGTT
LSS-Exon20-F	CAAGGCAGCCTGCTGGGGTGA
LSS-Exon20-R	CACCGGCTCACAGCTGAGTGT
LSS-Exon21-F	CTCACTGCAGCATTCCAGGGTT
LSS-Exon21-R	GTGGAAACAGCCATGCACGCT
LSS-Exon22-F	GCCAACAGCCAGGGCTCCAGTT
LSS-Exon22-R	GGTTGGAGCCCAAGACAGGGT

b

Gene	Primer (5'-3')
αA-R116C-For	TTCCCGTGAGTTCCACTGCCGCTACCGCCTGCCGTCGCTGC
αA-R116C-Rev	CGGCAGGCGGTAGCGGCAGTGGAACTCACGGG
αA-R116H-For	TTCCCGTGAGTTCCACCACCGCTACCGCCTGCCGTCGCCAC
αA-R116H-Rev	CGGCAGGCGGTAGCGGTGGTGGAACTCACGGG
αA-Y118D-For	GAGTTCCACCGCCGCGACCGCCTGCCGTCCAACTTACGAC
αA-Y118D-Rev	CGTTGGACGGCAGGCGGTCGCGGCGGTGGAACT
αB-R120G-For	CAGGGAGTTCCACGGGAAATACCGGATAGGGGG
αB-R120G-Rev	GGATCCGGTATTTCCCGTGGAACTCCCT
βB2-V187E-For	AGGTGCAGTCCGAGCGCCGTATGTGGAG
βB2-V187E-Rev	ATACGGCGCTCGGACTGCACCT
βB2-V187M-For	AGGTGCAGTCCATGCGCCGTATGTGATG
βB2-V187M-Rev	ATACGGCGCTCGGACTGCACCT
βB2-R188H-For	TGCAGTCCGTGCACCGTATCCCGCCAC
βB2-R188H-Rev	GGATACGGTGCACGGACTGCA
γC-G129C-For	CACGTGCTGGAGTGCTGCTGGGCTGC
γC-G129C-Rev	CAGCAGCACTCCAGCACGTG
γD-W43R-For	GTGGACAGCGGCTGCCGGATGCTCTATGAGCTGGCGG
γD-W43R-Rev	GCTCATAGAGCATCCGGCAGCCGCTGTCCAC

a, Primers used for PCR amplification and sequencing of each exon in the human *LSS* gene. **b**, Primers used in construction of crystallin mutants.

doi:10.1038/nature14468

T-cell exhaustion, co-stimulation and clinical outcome in autoimmunity and infection

Eoin F. McKinney[1,2], James C. Lee[1,2], David R. W. Jayne[1], Paul A. Lyons[1,2] & Kenneth G. C. Smith[1,2]

The clinical course of autoimmune and infectious disease varies greatly, even between individuals with the same condition. An understanding of the molecular basis for this heterogeneity could lead to significant improvements in both monitoring and treatment. During chronic infection the process of T-cell exhaustion inhibits the immune response, facilitating viral persistence[1]. Here we show that a transcriptional signature reflecting CD8 T-cell exhaustion is associated with poor clearance of chronic viral infection, but conversely predicts better prognosis in multiple autoimmune diseases. The development of CD8 T-cell exhaustion during chronic infection is driven both by persistence of antigen and by a lack of accessory 'help' signals. In autoimmunity, we find that where evidence of CD4 T-cell co-stimulation is pronounced, that of CD8 T-cell exhaustion is reduced. We can reproduce the exhaustion signature by modifying the balance of persistent stimulation of T-cell antigen receptors and specific CD2-induced co-stimulation provided to human CD8 T cells *in vitro*, suggesting that each process plays a role in dictating outcome in autoimmune disease. The 'non-exhausted' T-cell state driven by CD2-induced co-stimulation is reduced by signals through the exhaustion-associated inhibitory receptor PD-1, suggesting that induction of exhaustion may be a therapeutic strategy in autoimmune and inflammatory disease. Using expression of optimal surrogate markers of co-stimulation/exhaustion signatures in independent data sets, we confirm an association with good clinical outcome or response to therapy in infection (hepatitis C virus) and vaccination (yellow fever, malaria, influenza), but poor outcome in autoimmune and inflammatory disease (type 1 diabetes, anti-neutrophil cytoplasmic antibody-associated vasculitis, systemic lupus erythematosus, idiopathic pulmonary fibrosis and dengue haemorrhagic fever). Thus, T-cell exhaustion plays a central role in determining outcome in autoimmune disease and targeted manipulation of this process could lead to new therapeutic opportunities.

In a complex set of data such as the transcriptome, similar measurements may be grouped together by network analysis to form discrete modules that can highlight novel pathways contributing to the pathogenesis of complex diseases. We have previously shown that a CD8 T-cell transcriptional signature in patients with multiple immune-mediated diseases can predict a subsequent relapsing disease[2,3]. However, the biology underlying this observation was not clear. We therefore applied weighted gene co-expression network analysis (Extended Data Fig. 1) to the transcriptomes of purified CD4 and CD8 T cells isolated from a prospective cohort of 44 patients with anti-neutrophil cytoplasmic antibody-associated vasculitis (AAV) having active, untreated disease[2] (Supplementary Table 1) to further explore the mechanisms driving relapsing autoimmunity. Modules of genes (Fig. 1a, rows) were summarized as 'eigengene' profiles (Fig. 1b, f) that were correlated with clinical variables (Fig. 1a, i, columns) and visualized in the form of a heat map (Fig. 1a, i). Modules derived both from CD8 (Fig. 1a–d) and from CD4 (Fig. 1f–i) T-cell transcriptomes showed strong correlation with disease outcome but not activity, and were co-correlated (Fig. 1e) despite being mutually exclusive (Supplementary Table 2). A similar analysis using a cohort of 23 patients with systemic lupus erythematosus (SLE) also presenting with active, untreated disease (Supplementary Table 3)[2] identified analogous CD8 and CD4 T-cell expression modules (Extended Data Fig. 2) that again correlated with clinical outcome but not disease activity. By contrast a type 1 interferon (IFN) response signature was associated with disease activity but not with long-term outcome (Extended Data Fig. 2f), consistent with previous reports[4].

Next, we reasoned that genes within co-correlated modules in related cell types might inform the biology of relapsing disease. By selecting CD4 T-cell modules showing significant, strong correlation with relapse rate and performing network enrichment analysis, we identified a module corresponding to CD4 T-cell co-stimulation (Extended Data Figs 1f, g and 3a and Supplementary Tables 2 and 3). By way of validation, we repeated this analysis using an independent co-expression network algorithm that similarly demonstrated association between a CD4 co-stimulation module and clinical outcome (Supplementary Table 5). The independent association of modular signatures with clinical outcome (Fig. 1a, i) was confirmed using multiple linear regression modelling (Extended Data Fig. 3b–e) and was only apparent during active disease (Extended Data Fig. 3f and Supplementary Discussion).

During chronic viral infection, CD8 T-cell memory responses are exquisitely dependent on CD4 T-cell co-stimulation[5,6], which can lead to the resolution of chronic infection both in mice[1] and in humans[7]. When antigen persists in the absence of co-stimulation, CD8 T cells become 'exhausted'[1], a phenotype characterized by progressive loss of effector function, persistent high expression of inhibitory receptors and profound changes in gene expression, distinct from those seen in effector, memory or anergic T cells[8]. Although mice lacking inhibitory receptors have an increased incidence and severity of autoimmunity[9,10], a specific role for exhaustion in dictating the outcome of autoimmune responses has not been demonstrated.

We hypothesized that CD4 T-cell signals may be important in limiting exhaustion towards persistent self-antigen during autoreactive immunity, analogous to responses during persistent infection. We therefore used gene set enrichment analysis (GSEA[11]) to test for altered expression of transcriptional signatures reflecting T-cell exhaustion (and other T cell-related phenotypes) between subgroups of patients defined by the CD8 modular analysis, who go on to develop relapsing or quiescent autoimmunity (Fig. 2a). Using this approach, we observed that genes specifically downregulated in exhausted CD8 T cells during chronic murine lymphocytic choriomeningitis virus (LCMV) infection (but not altered in memory, naive or effector cells; Supplementary Table 6 (ref. 8)) were similarly downregulated in CD8 T cells from patients at low risk of subsequent relapse (Fig. 2b and Extended Data Figs 3g–i and 4).

During chronic murine LCMV infection, T-cell exhaustion is driven by coordinate upregulation of multiple co-inhibitory receptors[12] that signal synergistically to produce a state of generalized

[1]Department of Medicine, University of Cambridge School of Clinical Medicine, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK. [2]Cambridge Institute for Medical Research, University of Cambridge, Cambridge Biomedical Campus, Cambridge CB2 0XY, UK.

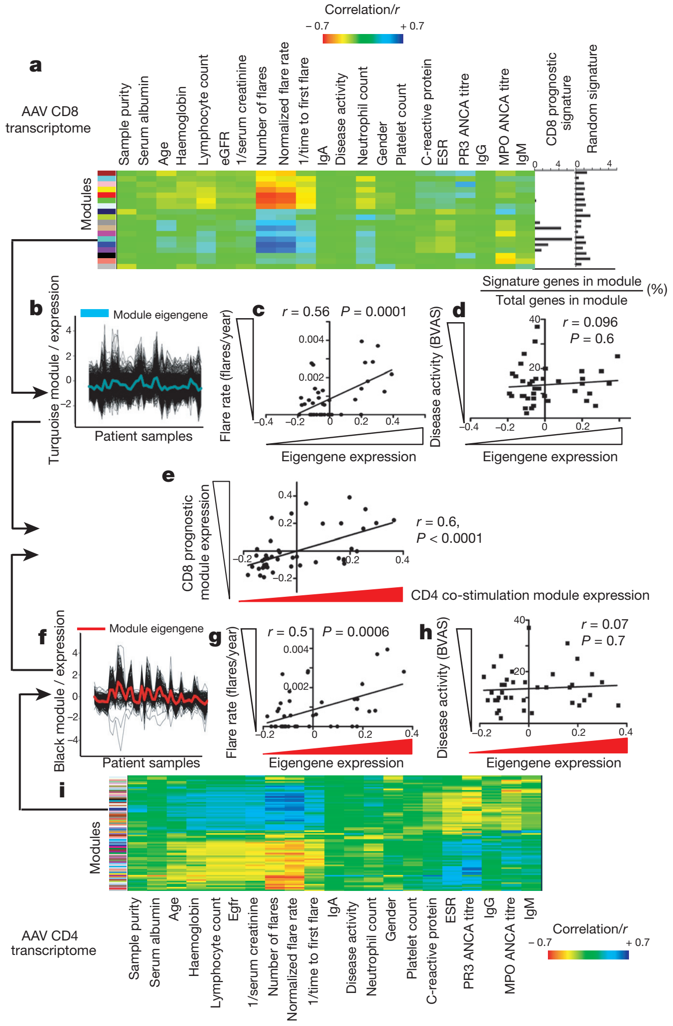


Figure 1 | Weighted gene co-expression network analysis of the T-cell transcriptome and its correlation with clinical phenotype in AAV. a, i, Heat maps illustrating the correlation of CD8 (**a**) and CD4 (**i**) co-expression modules (coloured blocks, *y* axis) with clinical traits in AAV ($n = 44$). Prognostic[2] and random signature overlap with modules shown (**a**, right) (overlap = signature genes/module genes, as a percentage). **b, f,** Linear plots illustrating turquoise (**b**) and black (**f**) modules and summary eigengenes; *y* = expression ($\log_2$(ratio)), *x* = samples. **c, d, g, h,** Scatter plots showing normalized flare rate (**c, g**) and disease activity (**d, h**, Birmingham Vasculitis Activity Score (BVAS), *y* axis) against CD8 turquoise (**c, d**) or CD4 black (**g, h**) module eigengene expression (*x* axis). **e,** Scatter plot showing correlation between CD4 T-cell black (*x* axis) and CD8 T-cell turquoise (*y* axis) module eigengenes. Pearson correlation, *r*, with *P* = two-tailed significance.

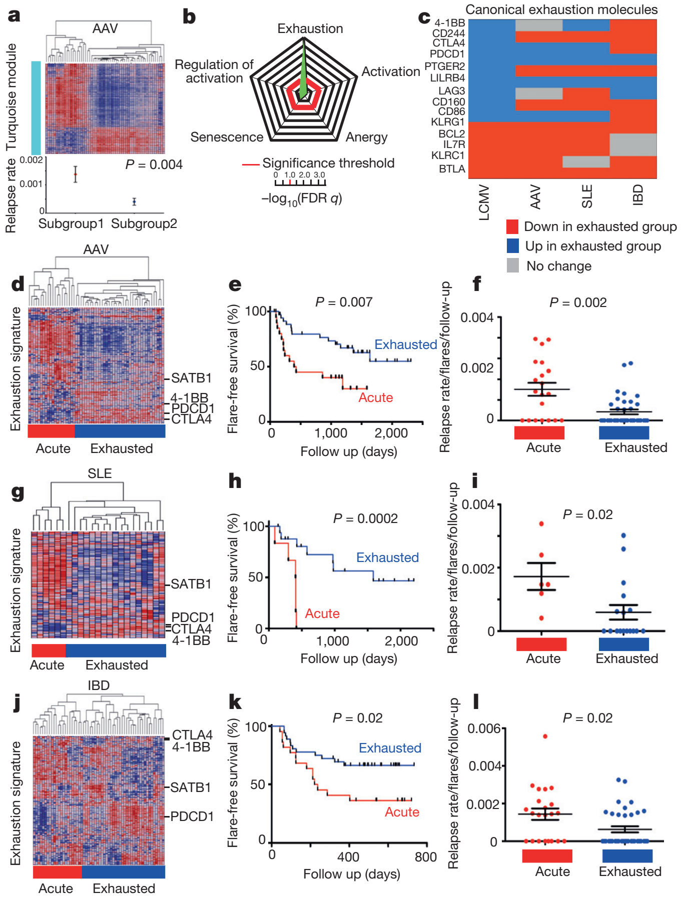


Figure 2 | A gene expression signature of CD8 T-cell exhaustion predicts contrasting outcomes in infection and autoimmune disease. a, Heat map showing hierarchical clustering of patients with AAV ($n = 44$) by expression of the turquoise module (Fig. 1b) with corresponding flare rates (flares per number of days follow-up, *y* axis). **b,** Wind rose plot showing GSEA significance (increasing from centre, $-\log_{10}$(false discovery rate (FDR) *q* value)) of CD8 T-cell signatures tested between prognostic subgroups defined in **a**. **c,** Heat map showing differential expression of exhaustion-associated co-inhibitory receptors between prognostic subgroups identified in **d, g, j**. Blue, up; red, down in exhausted; grey, no change (FDR $P < 0.05$). **d, g, j,** Heat maps showing hierarchical clustering of CD8 T-cell expression data from patients with AAV (**d**, $n = 58$), SLE (**g**, $n = 23$) and IBD (**j**, $n = 58$) using a murine CD8 exhaustion signature[8]. 'Exhausted' (blue) and 'non-exhausted' (red) subgroups of patients defined from the primary division of the cluster dendrogram. **e, h, k,** Kaplan–Meier curves showing censored flare-free survival and (**f, i, l**) scatter plots showing flare rate normalized against duration of follow-up for subgroups of patients defined in **d, g, j** for AAV (**e, f**), SLE (**h, i**) and IBD (**k, l**) cohorts. **e, h, k,** *P* = log-rank test. **a, f, i, l,** *P* = Mann–Whitney test.

immunosuppression[13]. In autoimmunity, these receptors were not coordinately upregulated as a group. Instead, patients with good prognosis from each disease were characterized by upregulation of a distinct subset of exhaustion-associated co-inhibitory receptors (Fig. 2c). Although a divergence from the murine LCMV model, T-cell exhaustion accompanied by expression of a limited subset of co-inhibitory receptors is similar to that described in intratumoural CD8 T cells[14], which are a target for checkpoint therapy (Extended Data Fig. 4i)[15,16].

To confirm whether exhaustion was associated with clinical outcome, we used the murine CD8 T-cell exhaustion signature (Supplementary Table 6 (ref. 8)) to perform unsupervised hierarchical clustering of three independent cohorts of patients with distinct diseases (AAV, Fig. 2d–f; SLE, Fig. 2g–i; inflammatory bowel disease (IBD), Fig. 2j–l). In each case this identified a subgroup of patients with both early (Fig. 2e, h, k) and recurrent (Fig. 2f, i, l) relapses. Whereas CD8 exhaustion was associated with poor outcome in viral infection, in every case it predicted favourable prognosis in autoimmune and inflammatory disease (Fig. 2d–l). Again, independent association with outcome was confirmed using multiple linear regression models (Extended Data Fig. 3d, e). Together, these data demonstrate that a transcriptional signature of relative CD8 T-cell exhaustion, similar to that determining outcome in chronic viral infection and cancer, is apparent during active, untreated disease in patients with favourable long-term outcome in multiple autoimmune and inflammatory diagnoses.

CD8 T-cell exhaustion is characterized by high expression of co-inhibitory receptors (such as PD-1 (ref. 12)) and low expression of nascent memory markers (such as interleukin (IL)-7R[17]) and is promoted both by the persistence of antigen[18] and by a lack of accessory co-stimulation[6]. To understand signals driving exhaustion and outcome in autoimmunity, we attempted to recreate the outcome-associated transcriptional signatures using variable T-cell antigen

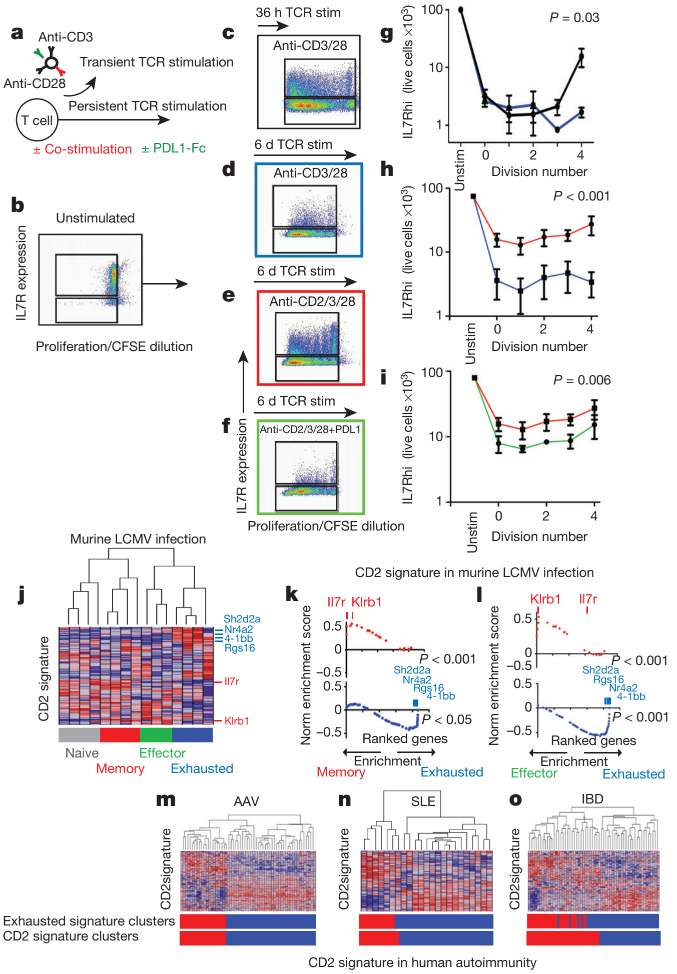


Figure 3 | T-cell co-stimulation with CD2 prevents development of an exhausted IL-7RloPD1hi phenotype. **a**, Schematic of the magnetic bead system providing variable TCR signal duration/co-stimulation during *in vitro* culture. **b–f**, Scatter plots illustrating IL-7R expression by cell division in unstimulated CD8 T cells (**b**) and following each of three different co-stimulation cultures (**c–f**), as indicated. **g–i**, Linear plots showing IL-7Rhi population resulting from (**g**) 36 h (black line) versus 6 d (blue line) anti-CD3/28 stimulation, (**h**) 6 d anti-CD2/3/28 (red line) versus 6 d anti-CD3/28 (blue line) and from 6 d anti-CD2/3/28 with (green line) and without (**i**, red line) Fc-PDL1. Although duration of stimulation varied, all cells were analysed after 6 days of culture. **j**, Heat map showing unsupervised hierarchical clustering of murine CD8 T-cell gene expression data[8] before (naive, grey), 8 days (effector, green) or 30 days (memory, red) after acute or more than 30 days (exhausted, blue) after chronic LCMV infection clustered by a CD2 response signature. **k**, **l**, Scatter plots showing GSEA enrichment for genes upregulated (red) and downregulated (blue) by CD2 in (**k**) memory versus exhausted and (**l**) effector versus exhausted CD8 T cells. **m–o**, Heat map showing unsupervised hierarchical clustering of AAV (**m**, $n = 58$), SLE (**n**, $n = 23$) and IBD (**o**, $n = 58$) CD8 T-cell expression data using the CD2 response signature. 'Exhausted' (blue) and 'non-exhausted' (red) subgroups were defined from the major division of the cluster dendrogram. Upper bar indicates comparison with subgroups of patients produced using the murine LCMV exhaustion signature (as shown in Fig. 2d, g, j). Enrichment by GSEA of CD2 signature in autoimmune subgroups with FDR $q < 0.1$.

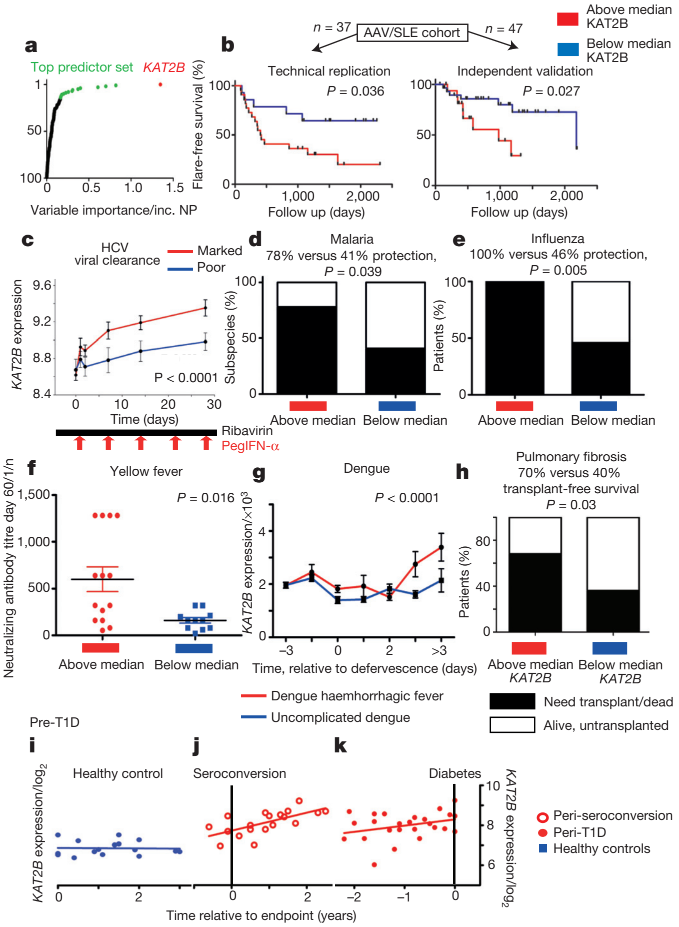


Figure 4 | A surrogate marker of CD4 co-stimulation in PBMC gene expression data correlates with clinical outcome in chronic viral infection, vaccination, infection and autoimmunity. **a**, Scatter plot showing the top 100 genes ranked by ability to identify CD4 T-cell co-stimulation subgroups in PBMC data; *x* axis, variable importance. **b**, Kaplan–Meier plots showing censored flare-free survival stratified by expression of *KAT2B* (red, above median; blue, below median) in patients with AAV and SLE ($n = 37$, training set) replicated on Affymetrix Gene 1.0 ST and in an independent cohort (test set, $n = 47$); P = log-rank test. **c**, Line and scatter plots showing serial *KAT2B* expression ($n = 54$) after therapy for chronic HCV infection giving a marked (red, $n = 28$) or poor response (blue, $n = 26$); P = two-way analysis of variance (ANOVA). **d**, Box plot showing post-vaccine malaria protection in a clinical trial ($n = 43$) stratified by *KAT2B* expression (red, above median; blue, below median); P = Fisher's exact test. **e**, Box plot showing percentage protection (black) in vaccinees ($n = 28$) after seasonal influenza vaccine stratified by *KAT2B* expression; P = Fisher's exact test. **f**, Scatter plot showing neutralizing antibody titre after YF-17D vaccination, stratified by *KAT2B* expression (**f**; red, above median *KAT2B*; blue, below median *KAT2B*); P = Mann–Whitney test. **g**, Line and scatter plot showing serial *KAT2B* expression throughout dengue infection ($n = 78$) stratified by progression to DHF (red, $n = 24$) or uncomplicated course (blue, uncomplicated dengue, $n = 54$). *x* axis, time (days) relative to defervescence. **h**, Box plot showing the percentage of patients with IPF ($n = 75$) progressing to transplantation/death (black) stratified by *KAT2B* expression (red, above median; blue, below median); P = Fisher's exact test. **i–k**, Scatter plots showing serial *KAT2B* expression in healthy age-, sex- and HLA-matched controls (**i**, blue) and in pre-T1D cases ($n = 5$, red), two of which seroconvert to islet-cell antibodies (**j**, black line) and three of which develop T1D (**k**, black line).

receptor (TCR) signal duration and co-stimulation of primary human cells *in vitro*. We stimulated purified human CD8 T cells using a magnetic bead conjugated with antibodies targeting co-stimulatory molecules (Fig. 3a) and measured expression of IL-7R and PD-1 expression (Fig. 3b–i and Extended Data Fig. 5a–g) as markers indicating an exhausted phenotype. Comparison between persistent (6 days) and transient (36 h) TCR stimulation showed that IL-7R expression returned on a proportion of cells after several divisions when the TCR stimulus was removed (Fig. 3c) but failed to do so if it persisted (Fig. 3d, g). We then systematically tested whether co-stimulatory molecules, identified from the CD4 T-cell network analysis described above (Fig. 1i and Extended Data Figs 3a and 5h–k), could overcome the effect of persistent TCR stimulation during *in vitro* differentiation. We found that specific co-stimulation with anti-CD2 (Fig. 3e, h), but not with other stimuli such as IFN-α or anti-CD40, resulted in maintained IL-7R expression, limited upregulation of PD-1 and enhanced cell survival (Fig. 3e and Extended Data Fig. 5).

While CD8 exhaustion is known to limit viral control during chronic infection, exhausted cells may be restored to useful function by blocking inhibitory signalling through PD-1 (ref. 19). Enhancing co-inhibitory signals is therefore a logical therapeutic strategy in autoimmune disease, aiming to facilitate exhaustion despite high levels of co-stimulation that would otherwise be predicted to result in an aggressive relapsing disease course. To test this concept, primary human CD8 T cells were co-stimulated during persistent TCR signalling as above (Fig. 3e) in the presence or absence of a bead-bound Fc-chimaeric version of the principal PD-1 ligand, PDL-1 (Fig. 3a, f). When added to CD2-co-stimulated CD8 T-cell cultures, increased PD-1/PDL-1 signalling suppressed differentiation of a non-exhausted IL-7R^{hi} subpopulation (Fig. 3f, i).

To define the phenotype of T-cell exhaustion more robustly, as small numbers of surface markers are insufficient, we analysed the transcriptome of CD8 T cells exposed to persistent stimulation with and without CD2 signalling (Supplementary Table 7). This CD2 response signature characterized exhausted cells but not effector or memory subsets (by GSEA; Fig. 3j–l). Consistent with this, patient clusters generated using the CD2 response signature recreated subgroups similar to those generated using the murine LCMV CD8 exhaustion signature (Figs 2d, g, j and 3m–o). Thus, CD2 signalling during persistent TCR stimulation of primary human CD8 T cells prevents the development of transcriptional changes characteristic of exhaustion, recreating transcriptional signatures associated with opposing outcomes in viral infection and autoimmunity.

To confirm that the transcriptional signatures reflected the development of functional exhaustion *in vitro*, we showed that cells appearing exhausted (IL-7R^{lo}PD-1^{hi}) also expressed markers of apoptotic resistance (BCL2^{lo}), characteristic cytokine patterns (IFN-γ^{lo}IL-10^{hi}) and showed diminished survival on re-stimulation (Extended Data Fig. 6a–e). There was no evidence of preferential accumulation of CD8 T-cell subsets after CD2-induced co-stimulation (Extended Data Fig. 6f–h). These data highlight the importance of CD2 signalling in limiting the development of CD8 T-cell exhaustion in the face of persistent TCR simulation, and provide a starting point for more sophisticated attempts to therapeutically exhaust an autoimmune response in a targeted fashion.

We next aimed to independently validate the association between the balance of CD4 co-stimulation and CD8 exhaustion with clinical outcome using published data sets. Most of these profile unseparated peripheral blood mononuclear cells (PBMCs), in which T-cell-intrinsic signatures are not readily apparent owing to the confounding influence of expression from other cell types[20]. We therefore used a classification algorithm (randomforests) to identify optimal surrogate markers of co-stimulation/exhaustion modules in PBMC from autoimmune patients taken concurrently with the T cells described above (Fig. 4a). As the CD8 exhaustion and CD4 co-stimulation signatures were themselves correlated (Extended Data Fig. 3g–i), it became easier to detect their combined signal in PBMC using surrogate markers (Fig. 4a and Extended Data Fig. 7). The top-ranked candidate, KAT2B, is a transcriptional co-activator known to mediate an anti-apoptotic effect under conditions of metabolic stress and to increase cellular resistance to cytotoxic compounds. These characteristics, along with its high expression in memory and T-follicular helper and natural killer (NK) cells (Extended Data Fig. 8), suggest that it may mark the development of a durable, persistent T-cell phenotype promoting long-lived responses in either infection or autoimmunity. The observed association was confirmed both by technical replication (using the same samples run on an independent array platform) and by independent validation (Fig. 4b).

To test whether similar associations may be apparent in multiple infectious and autoimmune diseases, we directly compared expression levels of *KAT2B* (and of the other top surrogate markers; Extended Data Fig. 9) between clinical subgroups defined within published studies for which PBMC expression and linked clinical outcome data were available. Where subgroups were not pre-specified, we compared clinical outcome in groups stratified as having either above- or below-median expression of *KAT2B* (Fig. 4c–k). Hierarchical clustering using all top surrogate markers gave similar stratification to that seen using *KAT2B* alone, while, as expected, the separation of subgroups of patients varied slightly in different clinical circumstances (Fig. 4c–k and Extended Data Fig. 9).

Combined IFN and ribavirin therapy may result in increased virus-specific T-cell responses in chronic hepatitis C virus (HCV) infection, although such eradication therapy is successful in only 50% of cases[21] and in some no change in endogenous immune response is observed[22]. In a cohort of patients with hepatitis C receiving combination therapy, *KAT2B* expression was progressively induced and showed significantly greater induction in those ultimately responding to therapy (Fig. 4c and Extended Data Fig. 10a). In a clinical trial of malaria vaccination[23], high *KAT2B* expression identified a subgroup with response rates of 78%, almost twice that of the low *KAT2B* expression group (Fig. 4d and Extended Data Fig. 10b–d). Moreover, response to vaccination for either influenza[24] (Fig. 4e and Extended Data Fig. 10e–f) or yellow fever[25] (Fig. 4f) could be predicted by stratifying recipients on the basis of their expression of *KAT2B* after exposure to vaccine. Dengue viral infection can result in a wide range of clinical manifestations ranging from asymptomatic infection or self-limiting fever (uncomplicated dengue) to dengue haemorrhagic fever (DHF). Consistent with our observations in autoimmunity, we observed that *KAT2B* expression was elevated in patients developing the excessive inflammatory response of DHF (Fig. 4g)[26].

We next asked whether surrogate detection of T-cell co-stimulation/exhaustion modules could predict progression of other autoimmune diseases. Idiopathic pulmonary fibrosis (IPF) is a progressive interstitial lung disease characterized by both autoantibodies and autoreactive CD4 T cells[27]. In a cohort of 75 patients with IPF[28], high expression of *KAT2B* predicted subsequent progression to transplantation or death (Fig. 4h). We also observed that PBMC *Kat2b* expression was elevated in the murine NOD model of type 1 diabetes (T1D)[29], with levels rising sharply during the T-cell initiation phase, long before the onset of diabetic hyperglycaemia (Extended Data Fig. 10g). In a cohort of samples taken prospectively from children at high risk of disease but before its onset[30], expression of *KAT2B* was seen to specifically and progressively rise (Fig. 4i–k) both in patients who progressed to T1D and in those who developed islet-cell autoantibodies.

We show that the balance between co-stimulatory and co-inhibitory signals that shape T-cell exhaustion coincides with opposite clinical outcomes during autoreactive and anti-viral immunity. This at once allows prediction of outcome during infection and autoimmunity, and creates the potential for targeted therapeutic exhaustion of an autoimmune response in those predicted to follow an aggressive disease course. That this association is apparent in multiple autoimmune and inflammatory diseases emphasizes the importance of signals shaping T-cell exhaustion in driving risk of relapse or recurrence (prognosis) rather than disease susceptibility (diagnosis) or immediate severity (disease activity), and suggests that targeted manipulation of these processes may lead to new treatment strategies that extend beyond the conditions discussed here.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 February 2014; accepted 10 April 2015.
Published online 29 June 2015.

1. Wherry, E. J. T cell exhaustion. *Nature Immunol.* **12,** 492–499 (2011).
2. McKinney, E. F. *et al.* A CD8$^+$ T cell transcription signature predicts prognosis in autoimmune disease. *Nature Med.* **16,** 586–591 (2010).

3. Lee, J. C. *et al.* Gene expression profiling of CD8$^+$ T cells predicts prognosis in patients with Crohn disease and ulcerative colitis. *J. Clin. Invest.* **121,** 4170–4179 (2011).
4. Baechler, E. C. *et al.* Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc. Natl Acad. Sci. USA* **100,** 2610–2615 (2003).
5. West, E. E. *et al.* Tight regulation of memory CD8$^+$ T cells limits their effectiveness during sustained high viral load. *Immunity* **35,** 285–298 (2011).
6. Aubert, R. D. *et al.* Antigen-specific CD4 T-cell help rescues exhausted CD8 T cells during chronic viral infection. *Proc. Natl Acad. Sci. USA* **108,** 21182–21187 (2011).
7. Urbani, S. *et al.* Outcome of acute hepatitis C is related to virus-specific CD4 function and maturation of antiviral memory CD8 responses. *Hepatology.* **44,** 126–139 (2006).
8. Wherry, E. J. *et al.* Molecular signature of CD8$^+$ T cell exhaustion during chronic viral infection. *Immunity* **27,** 670–684 (2007).
9. Rangachari, M. *et al.* Bat3 promotes T cell responses and autoimmunity by repressing Tim-3-mediated cell death and exhaustion. *Nature Med.* **18,** 1394–1400 (2012).
10. Francisco, L. M., Sage, P. T. & Sharpe, A. H. The PD-1 pathway in tolerance and autoimmunity. *Immunol. Rev.* **236,** 219–242 (2010).
11. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102,** 15545–15550 (2005).
12. Blackburn, S. D. *et al.* Coregulation of CD8$^+$ T cell exhaustion by multiple inhibitory receptors during chronic viral infection. *Nature Immunol.* **10,** 29–37 (2009).
13. Sevilla, N. *et al.* Immunosuppression and resultant viral persistence by specific viral targeting of dendritic cells. *J. Exp. Med.* **192,** 1249–1260 (2000).
14. Virgin, H. W., Wherry, E. J. & Ahmed, R. Redefining chronic viral infection. *Cell* **138,** 30–50 (2009).
15. Gubin, M. M. *et al.* Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature* **515,** 577–581 (2014).
16. Baitsch, L. *et al.* Exhaustion of tumor-specific CD8$^+$ T cells in metastases from melanoma patients. *J. Clin. Invest.* **121,** 2350–2360 (2011).
17. Lang, K. S. *et al.* Inverse correlation between IL-7 receptor expression and CD8 T cell exhaustion during persistent antigen stimulation. *Eur. J. Immunol.* **35,** 738–745 (2005).
18. Mueller, S. N. & Ahmed, R. High antigen levels are the cause of T cell exhaustion during chronic viral infection. *Proc. Natl Acad. Sci. USA* **106,** 8623–8628 (2009).
19. Barber, D. L. *et al.* Restoring function in exhausted CD8 T cells during chronic viral infection. *Nature* **439,** 682–687 (2006).
20. Lyons, P. A. *et al.* Microarray analysis of human leucocyte subsets: the advantages of positive selection and rapid purification. *BMC Genom.* **8,** 64 (2007).
21. Taylor, M. W. *et al.* Changes in gene expression during pegylated interferon and ribavirin therapy of chronic hepatitis C virus distinguish responders from nonresponders to antiviral therapy. *J. Virol.* **81,** 3391–3401 (2007).
22. Lauer, G. M. *et al.* Full-breadth analysis of CD8$^+$ T-cell responses in acute hepatitis C virus infection and early therapy. *J. Virol.* **79,** 12979–12988 (2005).
23. Vahey, M. T. *et al.* Expression of genes associated with immunoproteasome processing of major histocompatibility complex peptides is indicative of protection with adjuvanted RTS,S malaria vaccine. *J. Infect. Dis.* **201,** 580–589 (2010).
24. Nakaya, H. I. *et al.* Systems biology of vaccination for seasonal influenza in humans. *Nature Immunol.* **12,** 786–795 (2010).
25. Querec, T. D. *et al.* Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nature Immunol.* **10,** 116–125 (2009).
26. Hoang, L. T. *et al.* The early whole-blood transcriptional signature of dengue virus and features associated with progression to dengue shock syndrome in Vietnamese children and young adults. *J. Virol.* **84,** 12982–12994 (2010).
27. Shum, A. K. *et al.* BPIFB1 is a lung-specific autoantigen associated with interstitial lung disease. *Sci. Translat. Med.* **5,** 206ra139 (2013).
28. Herazo-Maya, J. D. *et al.* Peripheral blood mononuclear cell gene expression profiles predict poor outcome in idiopathic pulmonary fibrosis. *Sci. Translat. Med.* **5,** 205ra136 (2013).
29. Kodama, K. *et al.* Tissue- and age-specific changes in gene expression during disease induction and progression in NOD mice. *Clin. Immunol.* **129,** 195–201 (2008).
30. Elo, L. L. *et al.* Early suppression of immune response pathways characterizes children with prediabetes in genome-wide gene expression profiling. *J. Autoimmun.* **35,** 70–76 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by The National Institute of Health Research (NIHR), Cambridge Biomedical Research Centre and funded by the Wellcome Trust (project 094227/Z/10/Z and program grants 083650/Z/07/Z), a Lister Prize Fellowship (K.G.C.S) and the Lupus Research Institute (Distinguished Innovator Award, K.G.C.S). E.F.M. is a Wellcome–Beit Research Fellow supported by the Wellcome Trust and Beit Foundation (104064/Z/14/Z). The Cambridge Institute for Medical Research is in receipt of a Wellcome Trust Strategic Award (079895). We thank A. Kaser and J. Todd for reviewing the manuscript, staff of the NIHR Cambridge Biomedical Research Centre Cell Phenotyping hub and the patients who provided samples.

Author Contributions E.F.M. collected patients, analysed data and performed the *in vitro* experiments. E.F.M. wrote the manuscript with P.A.L. and K.G.C.S. E.F.M., K.G.C.S. and P.A.L. conceived the experiments and analysis. J.C.L. collected and processed samples from the IBD cohort and D.R.W.J. coordinated the review and follow-up of patients with AAV and SLE in the Cambridge Vasculitis clinic.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.F.M. (efm30@cam.ac.uk) or K.G.C.S. (kgcs2@cam.ac.uk).

METHODS

No statistical methods were used to predetermine sample size.

Ethical approval. Ethical approval for this study was obtained from the Cambridge Local Research Ethics Committee (reference numbers 04/023, 08/H0306/21, 08/H0308/176) and informed consent was obtained from all subjects enrolled.

Patients with AAV. Fifty-nine patients with AAV attending or referred to the specialist vasculitis unit at Addenbrooke's Hospital, Cambridge, UK, between July 2004 and May 2008 were enrolled into the present study. Active disease at presentation was defined by the BVAS[31] and the clinical impression that induction immunosuppression would be required. Prospective disease monitoring was undertaken monthly with serial BVAS disease scoring[31] and full biochemical, haematological and immunological profiling followed by treatment with an immunosuppressant and tapering dose steroid therapy (Supplementary Table 1). At each time-point of follow-up, disease activity was allocated into one of three categories defined as follows: (1) flare (at least one major or three minor BVAS criteria); (2) low-grade activity (no major and one or two minor BVAS criteria); (3) no activity (no major or minor BVAS criteria).

All disease flares were cross-checked against patient records to confirm clinical impression of disease activity and the need for intensified therapy as a result. Disease activity scoring was performed by a single investigator (E.F.M.) who was blinded to gene expression data at the time of scoring. Additional flares were defined in the absence of BVAS scoring if patients attended for emergency investigation (bronchoscopy, or specialist ophthalmological or ear/nose/throat surgical review) that confirmed evidence of active disease. To differentiate between discrete flares, clear improvement in disease activity was required in the form of an improvement in flare-related symptoms together with a reduction in BVAS score, a reduction in markers of inflammation (C-reactive protein, erythrocyte sedimentation rate) and a reduction in immunosuppressive therapy.

Patients with SLE. The SLE cohort comprised 23 patients attending or referred to the Addenbrooke's Hospital specialist vasculitis unit between July 2004 and May 2008 who met at least four American College of Rheumatology SLE criteria[32], presenting with active disease (defined below) and in whom immunosuppressive therapy was to be commenced or increased. After treatment with an immunosuppressant, patients were followed up monthly. Disease monitoring was undertaken with serial British Isles Lupus Assessment Group (BILAG) disease scoring[33] and full biochemical, haematological and immunological profiling (Supplementary Table 4).

A discrete disease flare required all three of the following prospectively defined criteria: (1) new BILAG score A or B in any system; (2) clinical impression of active disease by the reviewing physician; (3) the intention to increase immunosuppressive therapy as a result.

Additional flares were defined in the absence of BILAG scoring if patients were admitted directly to hospital as emergency cases for increased immunosuppressive therapy. To differentiate between disease flares, clear improvement in disease activity was required in the form of diminished flare-related symptoms together with a reduction in both BILAG score and immunosuppressive therapy.

Patients with IBD. Patients with active Crohn's disease and ulcerative colitis were recruited from a specialist IBD clinic at Addenbrooke's Hospital, before starting treatment. Diagnosis was made using standard endoscopic, histological and radiological criteria[34]. Patients who had already received immunomodulators or corticosteroids were excluded. Enrolled patients were managed conventionally using a step-up strategy[3].

Assessment of disease activity was in accordance with national and international guidelines and included consideration of symptoms, clinical signs and objective measures, including blood tests (C-reactive protein, erythrocyte sedimentation rate, haemoglobin concentration and serum albumin), stool markers (calprotectin) and mucosal assessment (by sigmoidoscopy or colonoscopy) where appropriate. Validated scoring tools were used as another means of assessing disease activity (Harvey–Bradshaw severity index[35] or simple clinical colitis activity index[36] for Crohn's disease and ulcerative colitis, respectively), although these were not used to guide treatment decisions. All clinicians were blinded to the microarray results.

For each disease, all patients were not included in all analyses as, for example, comparison of modular network analysis in related cell types required that samples passing quality control filtering were available for all cell types for all patients. Our previous publications have shown that the sample sizes used here are adequate to detect reproducible signatures correlating with clinical traits.

Follow-up analysis. Comparisons of outcome and associated clinical variables between subgroups were analysed using a Kaplan–Meier log-rank test and a non-parametric Mann–Whitney *U*-test or a χ^2 test as appropriate. Correction for multiple testing was applied using the Bonferroni method or FDR (Benjamini and Hochberg method) where appropriate as indicated.

Cell separation and RNA extraction. Venepuncture was performed at a similar time of day in all cases to minimize gene expression differences arising from circadian variation[37]. PBMCs, CD4 and CD8 T cells were isolated from 110 ml of whole blood by centrifugation over ficoll and, for T cells, by positive selection using magnetic beads as previously described[20]. The purity of separated cell subsets was determined by flow cytometry and included as a covariate in downstream correlation and network analyses (for example, Fig. 1a, i). Total RNA was extracted from each cell population using an RNeasy mini kit (Qiagen) with quality assessed using an Agilent BioAnalyser 2100 and RNA quantification performed using a NanoDrop ND-1000 spectrophotometer.

HsMediante 25k custom-spotted microarray. Total RNA (250 ng) was converted into double-stranded cDNA and labelled with Cy3- or Cy5-dCTP as previously described[20]. Appropriate Cy3- and Cy5-labelled samples were pooled and hybridized to custom-spotted oligonucleotide microarrays (HsMediante 25k) comprising probes representing 25,342 genes and control features[38]. All samples were hybridized in duplicate, using a dye-swap strategy, against a common reference RNA derived from pooled PBMC samples. After hybridization, arrays were washed and scanned on an Agilent G2565B scanner.

Affymetrix Human Gene ST microarray. Aliquots of total RNA (200 ng) were labelled using Ambion WT Sense Target labelling kit and hybridized to Human Gene 1.0 or 1.1 ST Arrays (Affymetrix) as described. After washing, arrays were scanned using a GS 3000 or Gene Titan scanner (Affymetrix) as appropriate.

Published data sets. Published data sets were accessed through either National Center for Biotechnology Information (NCBI) Gene Expression Omnibus or ArrayExpress, imported into R using the Bioconductor package GEOquery and analysed as described. Search criteria incorporated the name of individual diseases and were filtered to human data sets but not by platform used. Studies were only included if they met the following criteria. (1) Similar quality control filters as applied to the data produced in-house were satisfied (described below). (2) Samples were taken at an analogous time-point to those from which the co-stimulation and exhaustion signatures in autoimmunity were identified; that is, samples taken during active disease without concurrent immunosuppressive therapy. (3) Clinical outcome data were available.

It was not feasible to build a unified predictive model across all available data sets as they originated from different groups and were performed on mutually incompatible microarray platforms.

For the HCV data used in Fig. 4c, a marked response was defined as an HCV titre decrease greater than 3.5 $\log_{10}$(international units per millilitre (IU ml^{-1})) and a poor response as an HCV titre decrease less than 1.5 $\log_{10}$(IU ml^{-1}) by day 28 after commencing combined therapy with ribavirin and pegylated IFN-α. For the malaria vaccine trial used in Fig. 4d, 'protection' was defined as delayed or complete protection from subsequent confirmed *Plasmodium falciparum* infection after standardized exposure (five bites) compared with infectivity control subjects. For the influenza data used in Fig. 4e, protection was defined as at least one high response to at least one (of three) included strains. A high response was defined as at least a fourfold increase in haemagglutination inhibition titre at day 28 and a titre at least 1:40 as per US Food and Drug Administration guidelines.

All gene expression data used have been deposited in publicly available repositories (National Center for Biotechnology Information (NCBI) Gene Expression Omnibus and ArrayExpress): AAV, SLE (E-MTAB-2452, E-MTAB-157, E-MTAB-145), IBD (E-MTAB-331), LCMV (GSE9650), HCV (GSE7123), malaria vaccination (GSE18323), influenza vaccination (GSE29619), yellow fever vaccination (GSE13486), dengue fever (GSE25001), IPF (GSE28221), T1D (E-TABM-666), NOD (GSE21897), rheumatoid arthritis (GSE15258, GSE33377), *in vitro* CD8 stimulation (E-MTAB-3470).

Data preprocessing and quality control. For HsMediante 25k arrays, raw image data were extracted using Koadarray version 2.4 software (Koada Technology) and probes with a confidence score >0.3 in at least one channel were flagged as present. Extracted data were imported into R where log transformation and background subtraction were performed followed by within-array print-tip Loess normalization and between-array quantile and scale normalization using the Limma package[39] in Bioconductor[40]. Further analysis was then performed in R and only data demonstrating a strong negative correlation ($r^2 > 0.9$) between dye swap replicates were used in downstream analyses.

Affymetrix raw data (.cel) files were imported into R and subjected to variance stabilization normalization using the VSN package in BioConductor[41]. Quality control was performed using the Bioconductor package arrayQualityMetrics[42] with outlying samples removed from downstream analyses. Correction for batch variation was performed using the Bioconductor package ComBat[43] and batch structure was included as a covariate in downstream correlation analyses.

Clustering. Hierarchical clustering was performed using a Pearson correlation distance metric and average linkage analysis, performed either in Cluster with visualization in Treeview[44], using Genepattern[45], or directly in R using hclust in the statistics package.

Differential expression. Differentially expressed genes were identified using linear modelling and an empirical Bayes method[39] using an FDR threshold of 0.05 as indicated to determine significance.

Weighted gene co-expression network analysis. Highly correlated genes in immune cell subsets were identified and summarized with a modular eigengene profile using the weighted gene co-expression network analysis (WGCNA) bioconductor package in R[46]. Normalized, log-transformed expression data were variance-filtered using the inflexion point of a ranked list of median absolute deviation values for all probes. A soft thresholding power was chosen on the basis of the criterion of approximate scale-free topology[47]. Gene networks were constructed and modules identified from the resulting topological overlap matrix with a dissimilarity correlation threshold of 0.01 used to merge module boundaries and a specified minimum module size of $n = 30$. Modules were summarized as a network of modular eigengenes, which were then correlated with a matrix of clinical variables and the resulting correlation matrix visualized as a heat map (Extended Data Fig. 1). As each module by definition comprises highly correlated genes, their combined expression may be usefully summarized by eigengene profiles[48], effectively the first principal component of a given module (for example, Fig. 1b, f). A small number of eigengene profiles may therefore effectively 'summarize' the principle patterns within the cellular transcriptome with minimal loss of information. This dimensionality-reduction approach also facilitates correlation of modular eigengenes with clinical traits (for example, Fig. 1a, i). Significance of correlation between a given clinical trait and a modular eigengene was assessed using linear regression with Bonferroni adjustment to correct for multiple testing (Extended Data Fig. 1). Independent association of a given module eigengene or gene expression profile (for example, *KAT2B*) with clinical outcome was assessed using a multiple linear regression model. Significance of each term in the linear model was plotted against its regression coefficient, as a measure of the strength of association (the regression coefficient reflecting the change in clinical outcome per unit change in modular/gene expression), for example Extended Data Fig. 3b–e.

Overlap of signatures with modules derived from network analysis is shown to the right of selected module heat maps (Fig. 1a and Extended Data Fig. 2a, e, f) by the following formula to allow correction for variable module size: [(signature genes overlapping with module genes, n)/(genes in module, n)] × 100. The overlap of randomly selected signatures of equivalent size was used as a control and is shown adjacent to the above plots.

Hierarchical ordered partitioning and collapsing hybrid analysis. For validation purposes, highly correlated genes were independently partitioned into discrete modules using a second algorithm, hierarchical ordered partitioning and collapsing hybrid (HOPACH[49]) in R. This approach differs from WGCNA in that it does not rely on a user-specified correlation threshold to define module boundaries but rather aims to maximize homogeneity of modules. Normalized, log-transformed data were clustered using a hierarchical algorithm with modular boundaries defined by the median split silhouette, a measure of how well-matched a gene is to the other genes within its current cluster versus how well-matched it would be if it were moved to another cluster. On partitioning the data set into clusters, each cluster is reiteratively subdivided until the median split silhouette is maximized, thereby producing an optimal segregation into maximally discrete modules.

Knowledge-based network generation and pathway analysis. The biological relevance of gene groups composing modules identified by co-expression analysis was further investigated using the Ingenuity Pathway Analysis platform[50]. Six modules from the CD4 T-cell WGCNA analysis showed significant correlation with clinical outcome in AAV after correction for multiple testing (Bonferroni method; Supplementary Table 3). We applied network and pathway enrichment analysis to genes composing these modules to determine whether they may have any biological relevance. Briefly, for network analysis, genes from a specified target set of interest are progressively linked together on the basis of a measure of their interconnection, which is derived from described functional interactions. Additional highly interconnected genes that are absent from the target genes (open symbols) may be added to complete a network of arbitrary size (set at $n = 35$). Networks may be ranked by significance which reflects the probability of randomly generating a network of similar size from genes included in the full knowledge database containing at least as many target genes as in the network in question. For pathways analysis, the over-representation of canonical pathways (pre-defined, well-characterized metabolic and signalling pathways curated from extensive literature reviews) among a specified set of target genes is assessed, with significance determined by computing a Fisher's exact test with FDR correction for multiple testing.

GSEA. GSEA[11] was used to further assess whether specific biological pathways or signatures were significantly enriched between subgroups of patients identified by gene modules (as opposed to testing for enrichment of pathways within modules themselves as outlined in the previous section). GSEA determines whether an *a priori* defined 'set' of genes (such as a signature) shows statistically significant cumulative changes in gene expression between phenotypic subgroups (such as patients with relapsing or quiescent disease). In brief, all genes are ranked on the basis of their differential expression between two groups then an enrichment score is calculated for a given gene set on the basis of how often its members appear at the top or bottom of the ranked differential list. One thousand random permutations of the phenotypic subgroups were used to establish a null distribution of enrichment score against which a normalized enrichment score and FDR-corrected q values were calculated. GSEA was run with a focused subgroup of gene signatures (as in Figs 2b and 3k)[11] selected to test the null hypothesis that different CD8 T-cell phenotypes were not significantly enriched in subgroups of patients identified by modular analysis.

Selection of optimal PBMC-level biomarkers. Optimal surrogate markers facilitating identification of the CD4 T-cell co-stimulation/CD8 exhaustion signatures in PBMC-level data were determined using a random forests classification algorithm[51] (Fig. 4a). Although signatures apparent in purified T-cell transcriptome data correlate with clinical outcome, they cannot be similarly detected in data derived from PBMC owing to the confounding influence of expression from other cell types; nor can the same genes be used to predict outcome in PBMC[2,20]. However, as CD4 T-cell co-stimulation and CD8 T-cell exhaustion signatures themselves showed close correlation, we hypothesized that this would amplify the signal detectable in PBMC and that detection of the combined CD4/CD8 T-cell response might be feasible. The availability of both separated T-cell and PBMC data from the same patients at the same time facilitates a supervised search for surrogate markers of the co-stimulation/exhaustion signatures in PBMC. Expression data derived both from CD4 T cells and from PBMC were available for a cohort of $n = 37$ patients (AAV and SLE) after quality control and hybridization to the HsMediante 25k custom microarray platform and constituted a training cohort. Normalized, log-transformed expression data were analysed using the MLInterfaces Bioconductor package in R[52]. Using PBMC-level expression, data samples were classified into subgroups showing either high or low expression of the co-stimulation/exhaustion signature (as illustrated in Extended Data Fig. 5h, i) and probes were subsequently ranked using the variable importance metric on the basis of their ability to predict allocation to either group. The variable importance for a given gene reflects the change in accuracy of classification (percentage increase in MSE or increase in node purity) when that variable is randomly permuted. For a poorly predictive gene, random permutation of its values will minimally influence classification accuracy. Conversely, the most robust predictors will have a comparatively large effect on classification accuracy when randomly permuted. PBMC samples from a subset of $n = 37$ cases derived from the training cohort were labelled and hybridized on an alternative microarray platform (Affymetrix Gene 1.0 ST) as a technical validation set (Fig. 4b, left panel). PBMC samples from an independent cohort of $n = 47$ cases not included in the training cohort were labelled and hybridized to the Affymetrix Gene 1.0 ST platform as an independent test set (Fig. 4b, right panel). For both technical validation and independent test sets, expression of the optimal biomarker identified in Fig. 4a (*KAT2B*) was used to bisect the cohort relative to the median expression and clinical outcome was compared in patients with *KAT2Bhi* and *KAT2Blo*.

Linear models. Linear modelling was performed in R using the statistics package. This took the form of

$$fit < -lm(y \sim x1 + x2 + x3,\ data = mydata),$$

where y (the response variable) was selected as normalized flare rate (flares per number of days follow-up, and $x1$–xn (the test variables) were selected to include measures of disease activity (both clinical scores and laboratory markers of inflammation), quantification of circulating leucocyte subsets (lymphocytes, neutrophils) and concurrent measurements of autoantibody titre where relevant. Test variables also included a biomarker profile (for example, exhaustion signature or *KAT2B* expression). The significance and magnitude (regression coefficient, reflecting change in response variable (flares per number of days follow-up) per unit change in each test variable included) were extracted and plotted against each other (for example, Extended Data Fig. 3b–e). Not all clinical or laboratory measures were relevant comparisons in each case and therefore were not all included in every model generated.

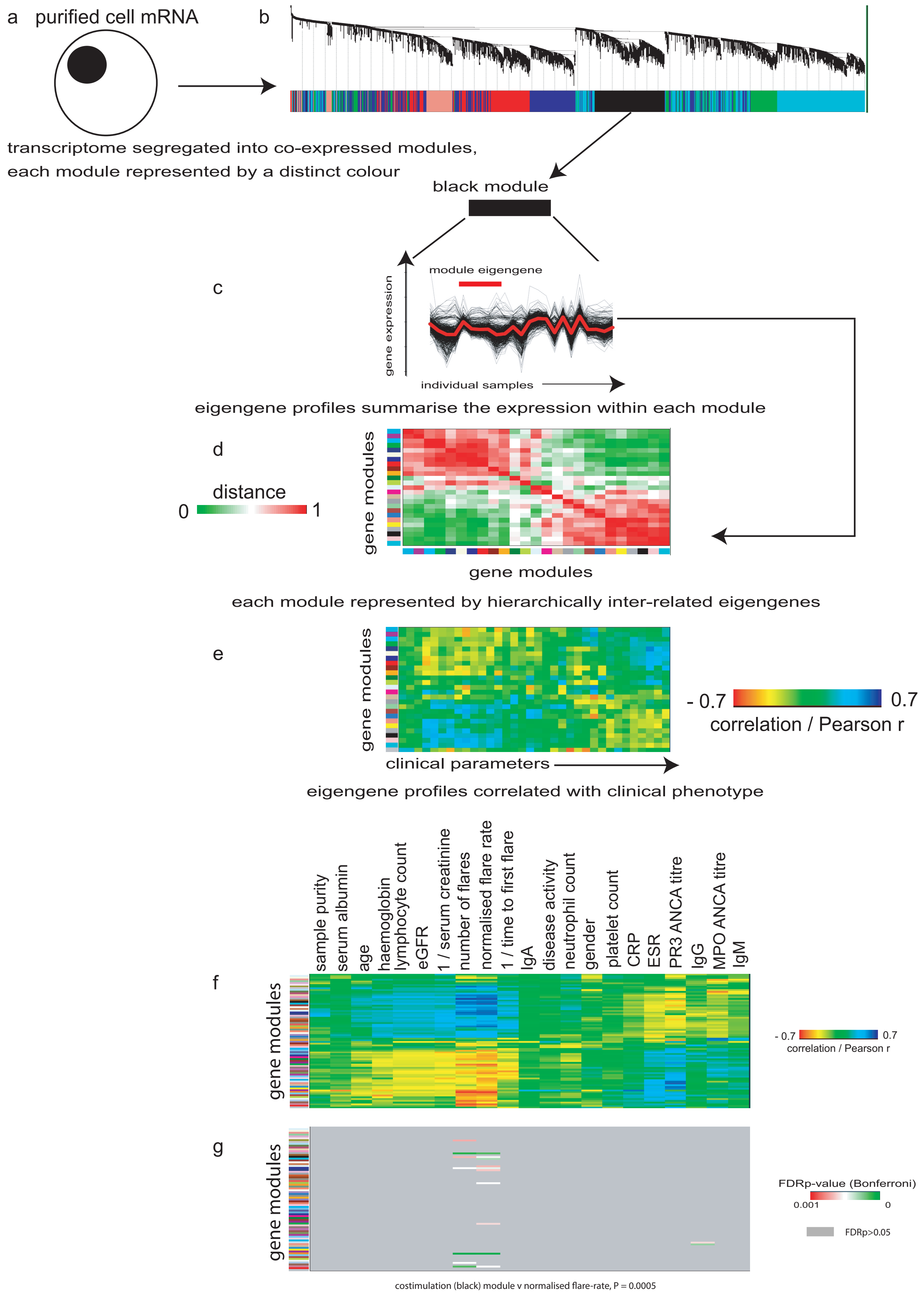
T-cell culture. Primary human CD8 T cells were separated from leucocyte cones obtained from NHS Blood and Transplant (Addenbrooke's Hospital) by centrifugation over ficoll and positive selection using magnetic beads as previously described[20]. The purity of separated cell subsets was determined by three-colour flow cytometry. Purified T cells were labelled with 10 μM CFSE (Invitrogen) and re-suspended in complete RPMI 1640 (Sigma Aldrich) in the presence of 10% FCS. Purified CD8+ T cells (>95%) were then stimulated in sterile, 96-well U-bottomed culture plates (Greiner) using an 'artificial APC' consisting of MACS iBead particles (1:2 bead:cell ratio, Miltenyi) or DynaBead particles (Invitrogen) conjugated to either CD3/CD28 or CD2/CD3/CD28 as indicated in the presence of IL-2 (10 ng ml^{-1}, Gibco Life Technologies) for 6 days. The magnetic iBead construct was removed after 36 h in some instances as indicated. In some experiments, additional co-stimulation was provided by the addition of either IFN-α (10 ng ml^{-1}, Abcam) or by additional conjugation of recombinant Human PD-L1 Fc Chimera (Life Technologies, 1 μg ml^{-1}) or anti-CD40 antibody (50 ng ml^{-1}, Abcam) as indicated. The nature of co-stimulatory signals tested was based upon the results of the network analysis of CD4 T-cell modules described above (Supplementary Table 2).

For re-stimulation experiments, cells were harvested on day 6 after stimulation and sorted into IL-7R^{hi} and IL-7R^{lo} populations (Extended Data Fig. 6d) using a FACSAriaIII cell sorter (BD Biosciences) with live/dead discrimination performed using an AquaFluorescent amine-reactive dye (Invitrogen). Cell numbers were normalized and were re-suspended in complete RPMI 1640 (2 × 10^4 per millilitre, Sigma-Aldrich) and 'rested' in a sterile, U-bottomed culture plate (Greiner) for 6 days (37°C, 5% CO$_2$) before being re-stimulated (anti-CD2/3/28 1:2 bead:cell ratio, Miltenyi MACSiBead) for a further 6 days in the presence of IL-2 (10 ng ml^{-1}, Gibco Life Technologies).

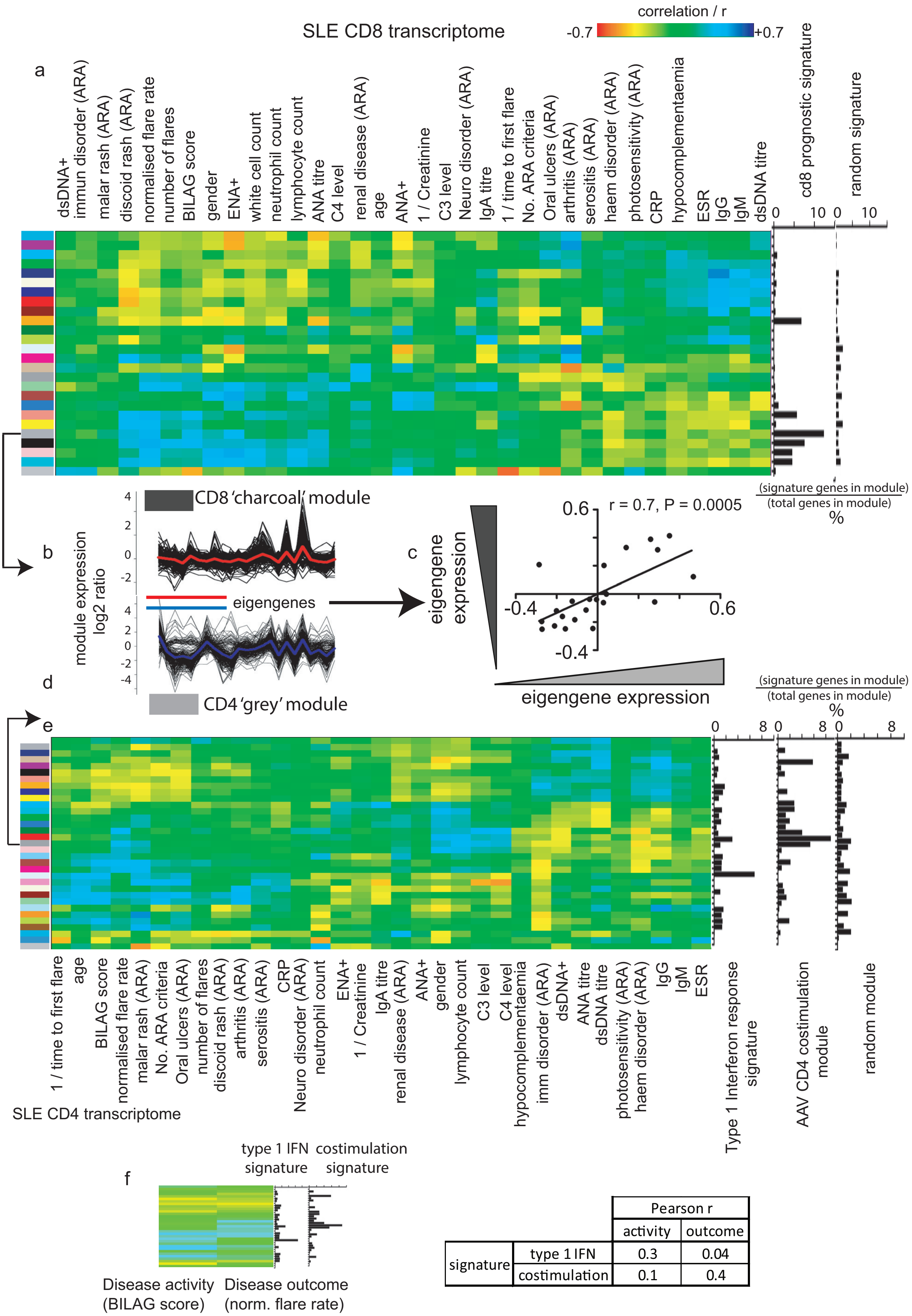
Note that, as described in Extended Data Fig. 6g, human memory CD8 T-cell subsets do not equivalently respond to the stimulation conditions described above. As primary whole human CD8 T cells are composed of highly variable proportions of memory subsets and whole CD8 T cells were stimulated, it was necessary to perform paired tests of significance when comparing resulting T-cell subsets and transcriptional profiles.

Flow cytometry. Immunophenotyping was performed using an LSR Fortessa analyser (BD Biosciences), and data were analysed using FlowJo software (Tree Star). Reactions were standardized with multicolour calibration particles (BD Biosciences) with saturating concentrations of the following antibodies: AquaFluorescent Live/Dead (Invitrogen), IL-7Rα AF647 (BD biosciences, clone HIL-7R-M21), PDCD1 APC (eBioscience, clone MIH4). For intracellular staining, cells were fixed and permeabilized using a transcription factor staining buffer set (eBioscience) and before staining with saturating concentrations of antibody against BCL2 (BD Biosciences, clone 100).

31. Stone, J. H. *et al.* A disease-specific activity index for Wegener's granulomatosis: modification of the Birmingham Vasculitis Activity Score. International Network for the Study of the Systemic Vasculitides (INSSYS). *Arthritis Rheum.* **44,** 912–920 (2001).
32. Tan, E. M. *et al.* The 1982 revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum.* **25,** 1271–1277 (1982).
33. Isenberg, D. A. *et al.* BILAG 2004. Development and initial validation of an updated version of the British Isles Lupus Assessment Group's disease activity index for patients with systemic lupus erythematosus. *Rheumatology* **44,** 902–906 (2005).
34. Silverberg, M. S. *et al.* Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: report of a Working Party of the 2005 Montreal World Congress of Gastroenterology. *Can. J. Gastroenterol.* **19,** 5A–36A (2005).
35. Harvey, R. F. & Bradshaw, M. J. Measuring Crohn's disease activity. *Lancet* **i,** 1134–1135 (1980).
36. Walmsley, R. S., Ayres, R. C., Pounder, R. E. & Allan, R. N. A simple clinical colitis activity index. *Gut* **43,** 29–32 (1998).
37. Whitney, A. R. *et al.* Individuality and variation in gene expression patterns in human blood. *Proc. Natl Acad. Sci. USA* **100,** 1896–1901 (2003).
38. Le Brigand, K. *et al.* An open-access long oligonucleotide microarray resource for analysis of the human and mouse transcriptomes. *Nucleic Acids Res.* **34,** e87 (2006).
39. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3,** 3 (2004).
40. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).
41. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18** (Suppl. 1), S96–S104 (2002).
42. Kauffmann, A., Gentleman, R. & Huber, W. arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics* **25,** 415–416 (2009).
43. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2007).
44. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95,** 14863–14868 (1998).
45. Renshaw, B. R. *et al.* Humoral immune responses in CD40 ligand-deficient mice. *J. Exp. Med.* **180,** 1889–1900 (1994).
46. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformat.* **9,** 559 (2008).
47. Barabasi, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286,** 509–512 (1999).
48. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst. Biol.* **1,** 54 (2007).
49. van der Laan, M. J. & Pollard, K. S. A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *J. Stat. Plann. Inf.* **117,** 275–303 (2002).
50. Ingenuity Pathway Analysis (Ingenuity Systems, 2003).
51. Breiman, L. Random forests. *Machine Learn. J.* **45,** 5–32 (2001).
52. Gentleman, R. *et al. Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (Springer, 2005).
53. Cramp, M. E. *et al.* Hepatitis C virus-specific T-cell reactivity during interferon and ribavirin treatment in chronic hepatitis C. *Gastroenterology* **118,** 346–355 (2000).
54. Bienkowska, J. R. *et al.* Convergent random forest predictor: methodology for predicting drug response from genome-scale data applied to anti-TNF response. *Genomics* **94,** 423–432 (2009).
55. Toonen, E. J. *et al.* Validation study of existing gene expression signatures for anti-TNF treatment in patients with rheumatoid arthritis. *PLoS ONE* **7,** e33199 (2012).

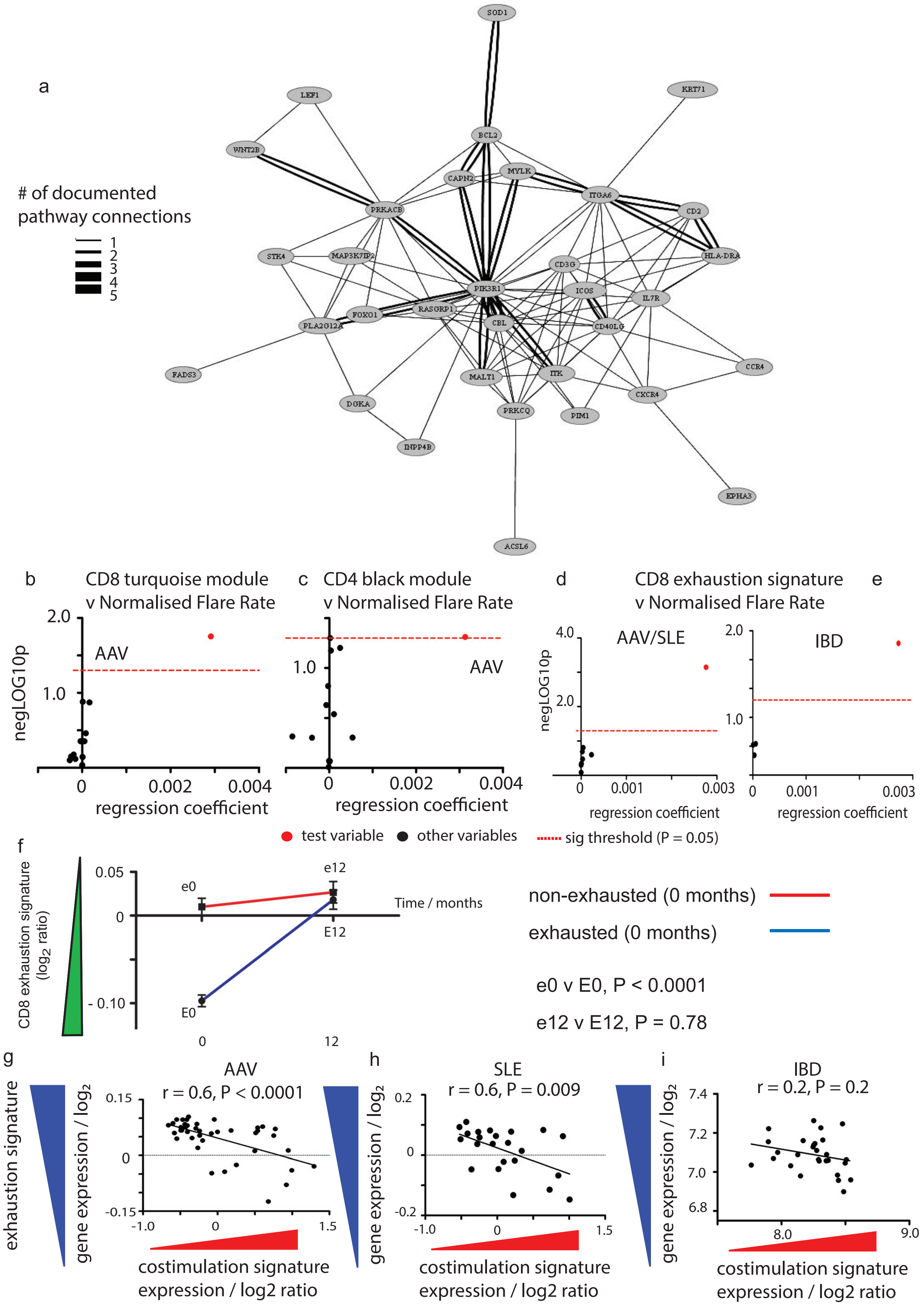

a
purified cell mRNA
b
transcriptome segregated into co-expressed modules,
each module represented by a distinct colour
black module
c
module eigengene
gene expression
individual samples
eigengene profiles summarise the expression within each module
d
distance
0
1
gene modules
gene modules
each module represented by hierarchically inter-related eigengenes
e
gene modules
- 0.7
0.7
correlation / Pearson r
clinical parameters
eigengene profiles correlated with clinical phenotype
f
sample purity
serum albumin
age
haemoglobin
lymphocyte count
eGFR
1 / serum creatinine
number of flares
normalised flare rate
1 / time to first flare
IgA
disease activity
neutrophil count
gender
platelet count
CRP
ESR
PR3 ANCA titre
IgG
MPO ANCA titre
IgM
gene modules
- 0.7
0.7
correlation / Pearson r
g
gene modules
FDRp-value (Bonferroni)
0.001
0
FDRp>0.05
costimulation (black) module v normalised flare-rate, P = 0.0005

Extended Data Figure 1 | Overview of weighted gene co-expression analysis. **a**, Messenger RNA derived from purified leucocyte subsets sampled during active, untreated autoimmune disease is labelled and hybridized to a microarray platform (both HsMediante 25k and Affymetrix Gene 1.0 ST used here). Genes are then combined into modules (**b**, coloured blocks) based on the similarity of their expression profile in all samples. **c**, Detail for the 'black' module. Each horizontal black line represents expression of a single gene within the given module; *y* axis, gene expression; *x* axis, patient samples; red bar, eigengene profile which effectively summarizes the expression of all genes constituting the black module. **d**, Each modular profile is related to all others in a hierarchy that can itself be visualized by plotting correlation of all module eigengenes, such as in the heat map shown here. Coloured blocks represent individual modules, defined as in **a**. Modules are aligned in identical order on *x* and *y* axes with heat-map colour representing the correlation between each. Note that the diagonal (top left to bottom right) therefore represents the correlation of each eigengene profile with itself, and is always 1. Distance metric is the Euclidean distance. **e**, As each module is summarized by a representative eigengene profile, each may then be correlated against a range of clinical variables, allowing visualization of how the transcriptome relates to clinical variables, again in the form of a correlation heat map. Pearson correlation, *r*. **f**, Heat map showing gene expression modules (*y* axis) correlated against clinical variables (*x* axis) for the CD4 transcriptome in AAV. Pearson correlation, *r*. **g**, Heat map illustrating significance of correlations identified in **f**. *P* value threshold at Bonferroni-corrected $P < 0.05$. Colour bar indicates actual *P* value of correlations deemed significant; grey shading, corrected $P > 0.05$. Significance for co-stimulation (black) module from Fig. 1 is also shown ($P = 0.0005$).

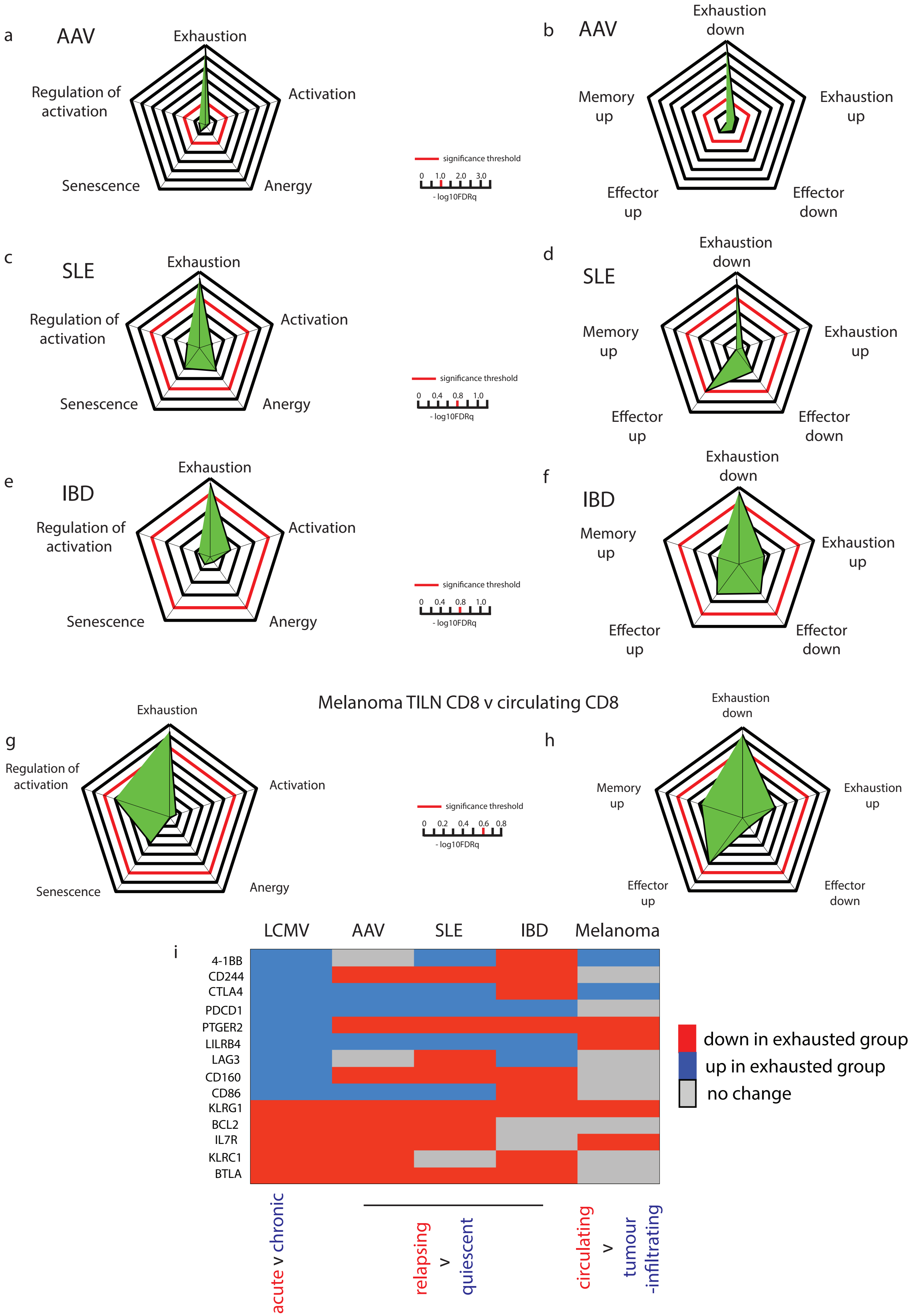


		Pearson r	
		activity	outcome
signature	type 1 IFN	0.3	0.04
	costimulation	0.1	0.4

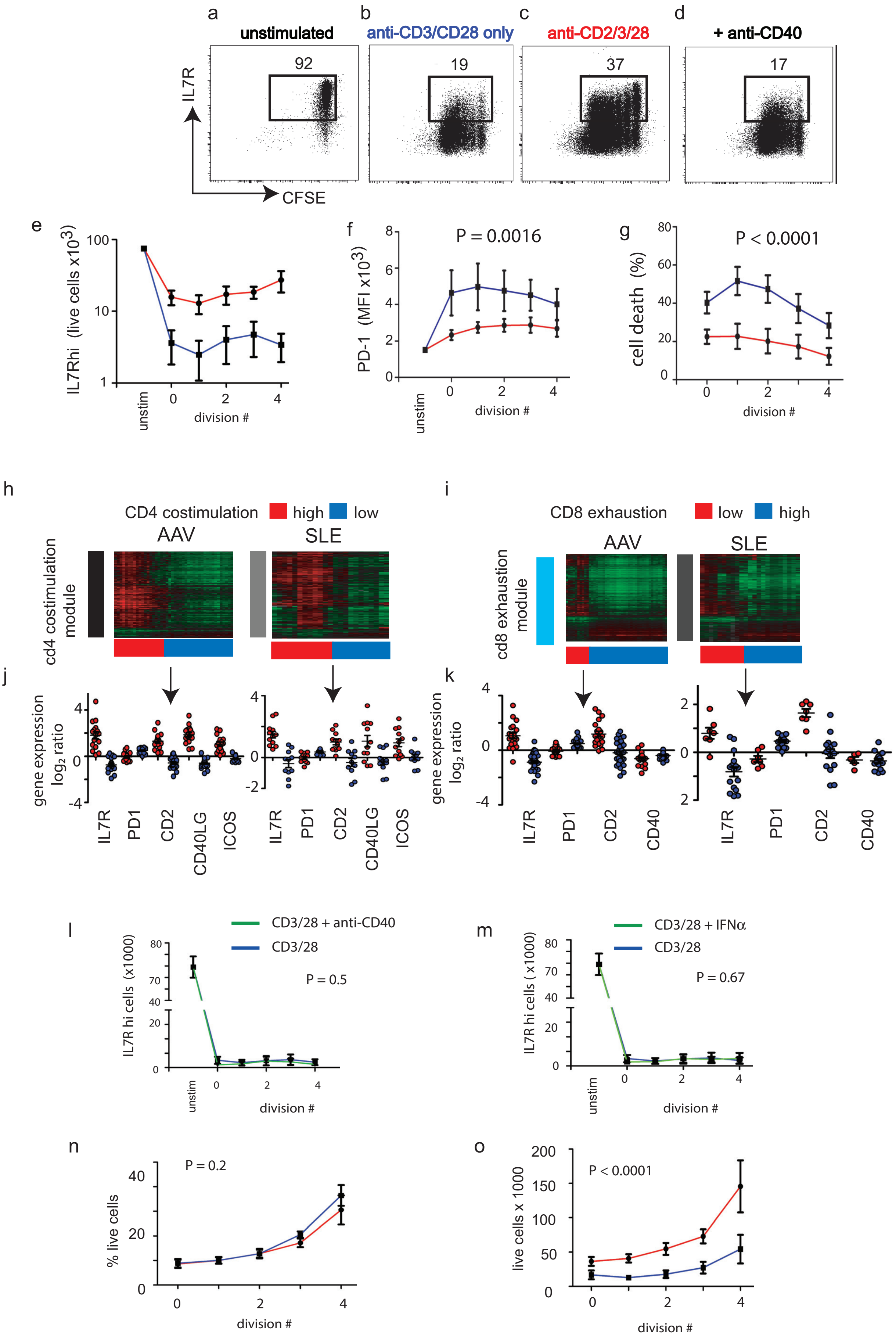
Extended Data Figure 2 | Weighted gene co-expression network analysis of the T-cell transcriptome and its correlation with clinical phenotype in SLE. **a**, **e**, Heat maps illustrating the correlation of co-expression modules (coloured blocks, *y* axis) derived from the CD8 (**a**) and CD4 (**e**) transcriptomes of 23 SLE patients with clinical traits (*x* axis). Overlap of the previously described prognostic signature with co-expression modules, along with the distribution of a random signature of equivalent size, shown to the right of **a** (overlap = signature genes/module genes, as a percentage). Overlap of the CD4 T-cell co-stimulation 'black' module (defined in Fig. 1) shown to the right of **e**, with a randomly derived module and a type 1 IFN response signature previously shown to associate with active SLE[4]. Overlap shown as percentage representation of the signature within each module. **b**, **d**, Linear plots illustrating the 'charcoal' (**b**) and 'grey' (**d**) modules in detail; *y* axis, gene expression; *x* axis, individual patients; coloured lines (red, blue), module eigengenes. **c**, Correlation of SLE CD4 T-cell co-stimulation module eigengene (*x* axis, blue) against SLE CD8 T-cell prognostic signature (*y* axis, red). Pearson correlation, *r*, with *P* = two-tailed significance. **f**, Expanded detail from **e**, illustrating that modules corresponding to type 1 IFN response and co-stimulation signatures correlate with disease activity and outcome respectively but not vice versa.


a
# of documented pathway connections
1
2
3
4
5
SOD1
LEF1
KRT71
BCL2
WNT2B
CAPN2
MYLK
ITGA6
PRKACB
CD2
STK4
MAP3K7IP2
HLA-DRA
CD3G
PIK3R1
ICOS
IL7R
RASGRP1
FOXO1
PLA2G12A
CBL
CD40LG
FADS3
CCR4
MALT1
ITK
DGKA
CXCR4
PRKCQ
PIM1
INPP4B
EPHA3
ACSL6
b
CD8 turquoise module v Normalised Flare Rate
negLOG10p
2.0
1.0
AAV
0
0.002
0.004
regression coefficient
c
CD4 black module v Normalised Flare Rate
1.0
AAV
0
0.002
0.004
regression coefficient
d
CD8 exhaustion signature v Normalised Flare Rate
negLOG10p
4.0
2.0
AAV/SLE
0
0.001
0.003
regression coefficient
e
2.0
1.0
IBD
0
0.001
0.003
regression coefficient
● test variable
● other variables
sig threshold (P = 0.05)
f
CD8 exhaustion signature (log$_2$ ratio)
0.05
0
- 0.10
e0
e12
E12
E0
Time / months
0
12
non-exhausted (0 months)
exhausted (0 months)
e0 v E0, P < 0.0001
e12 v E12, P = 0.78
g
exhaustion signature
gene expression / log$_2$
AAV
r = 0.6, P < 0.0001
0.15
0
-0.15
-1.0
0
1.5
costimulation signature expression / log2 ratio
h
gene expression / log$_2$
SLE
r = 0.6, P = 0.009
0.2
0
-0.2
-1.0
0
1.5
costimulation signature expression / log2 ratio
i
gene expression / log$_2$
IBD
r = 0.2, P = 0.2
7.4
7.2
7.0
6.8
8.0
9.0
costimulation signature expression / log2 ratio

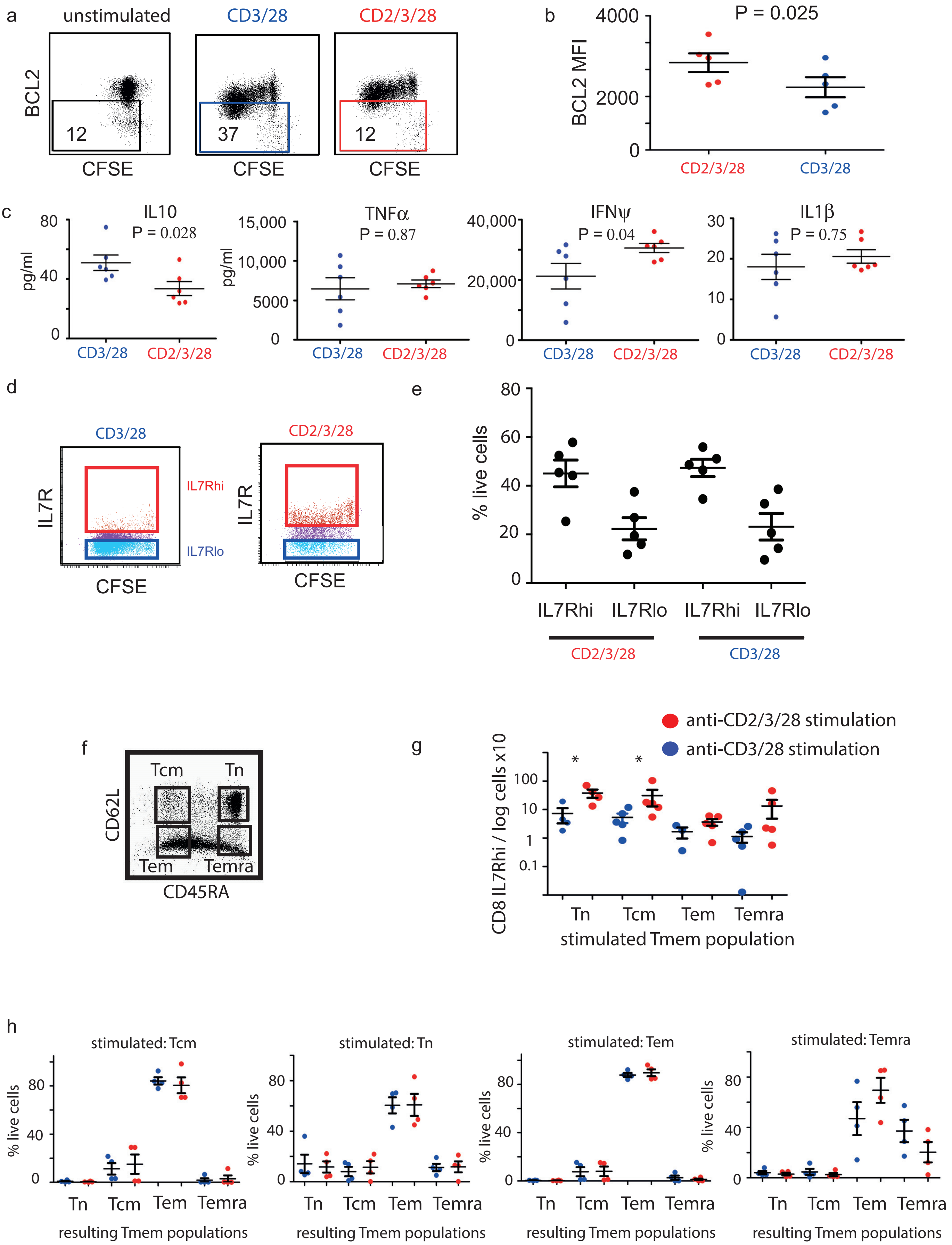
Extended Data Figure 3 | Identification and validation of genes involved in CD4 co-stimulation that correlate with clinical outcome, and how that relationship changes after treatment. **a**, A knowledge-based network analysis of 336 probes composing the 'black' expression module (Fig. 1e) identifies a network of co-stimulation signalling (Supplementary Table 3). Individual genes are shown in circles with the 'strength' of their connections indicated by the weight of the black bar linking them. Pathways of TCR signalling, inducible T-cell co-stimulator and its ligand (ICOS–ICOSL) signalling and CD28 signalling are all significantly enriched in this module (FDR $P < 0.05$). **b**–**e**, Scatter plots showing the outcome of multiple linear regression models testing the association of four signatures (red symbols) as indicated, directly compared with clinical markers of disease activity (black symbols); *x* axis, magnitude of association (regression coefficient, change in normalized flare rate (flares per number of days follow-up) per unit change in each variable tested); *y* axis, significance of association in multiple regression model; *P*, significance threshold (dashed red line, $P = 0.05$). **b**, CD8 turquoise module eigengene in AAV, (**c**) CD4 co-stimulation (black) module eigengene in AAV, (**d**, **e**) CD8 exhaustion signature (Supplementary Table 6) in AAV/SLE (**d**) and IBD (**e**). Clinical variables incorporated vary owing to differing relevance in each case but include some of the following: disease activity score (BVAS/BILAG/CDAI/Harvey–Bradshaw score), C-reactive protein, autoantibody titre (PR3/MPO, dsDNA), lymphocyte count, neutrophil count, platelet count, IgG, IgA, IgM, erythrocyte sedimentation rate, age. **f**, Line plot showing mean expression of a CD8 T-cell exhaustion signature in 38 patients with AAV measured at presentation during active, untreated disease (t_0) and 12 months later when disease activity was quiescent and patients were on maintenance immunosuppressive therapy (t_{12}). Patients are grouped into those falling above (red) and below (blue) median expression of the exhaustion signature eigengene at entry. $P =$ Mann–Whitney test comparing t_{12} and t_0 values. The difference between the groups that is easily apparent at enrolment with active, untreated disease (t_0) is no longer apparent when disease is treated and quiescent 12 months later (t_{12}). **g**–**i**, Scatter plots showing inverse correlation between individual eigenvalues of the CD4 co-stimulation signature (*x* axis, red) and the CD8 exhaustion signature (*y* axis, blue) defined as in Fig. 2, for AAV (**g**), SLE (**h**) and IBD (**i**) cohorts. Pearson correlation, r^2, two-tailed significance.


a
AAV
Exhaustion
Activation
Regulation of activation
Senescence
Anergy
significance threshold
0 1.0 2.0 3.0
- log10FDRq
b
AAV
Exhaustion down
Exhaustion up
Memory up
Effector up
Effector down
c
SLE
Exhaustion
Activation
Regulation of activation
Senescence
Anergy
significance threshold
0 0.4 0.8 1.0
- log10FDRq
d
SLE
Exhaustion down
Exhaustion up
Memory up
Effector up
Effector down
e
IBD
Exhaustion
Activation
Regulation of activation
Senescence
Anergy
significance threshold
0 0.4 0.8 1.0
- log10FDRq
f
IBD
Exhaustion down
Exhaustion up
Memory up
Effector up
Effector down
Melanoma TILN CD8 v circulating CD8
g
Exhaustion
Activation
Regulation of activation
Senescence
Anergy
significance threshold
0 0.2 0.4 0.6 0.8
- log10FDRq
h
Exhaustion down
Exhaustion up
Memory up
Effector up
Effector down
i
LCMV
AAV
SLE
IBD
Melanoma
4-1BB
CD244
CTLA4
PDCD1
PTGER2
LILRB4
LAG3
CD160
CD86
KLRG1
BCL2
IL7R
KLRC1
BTLA
🟥 down in exhausted group
🟦 up in exhausted group
⬜ no change
acute v chronic
relapsing v quiescent
circulating v tumour -infiltrating

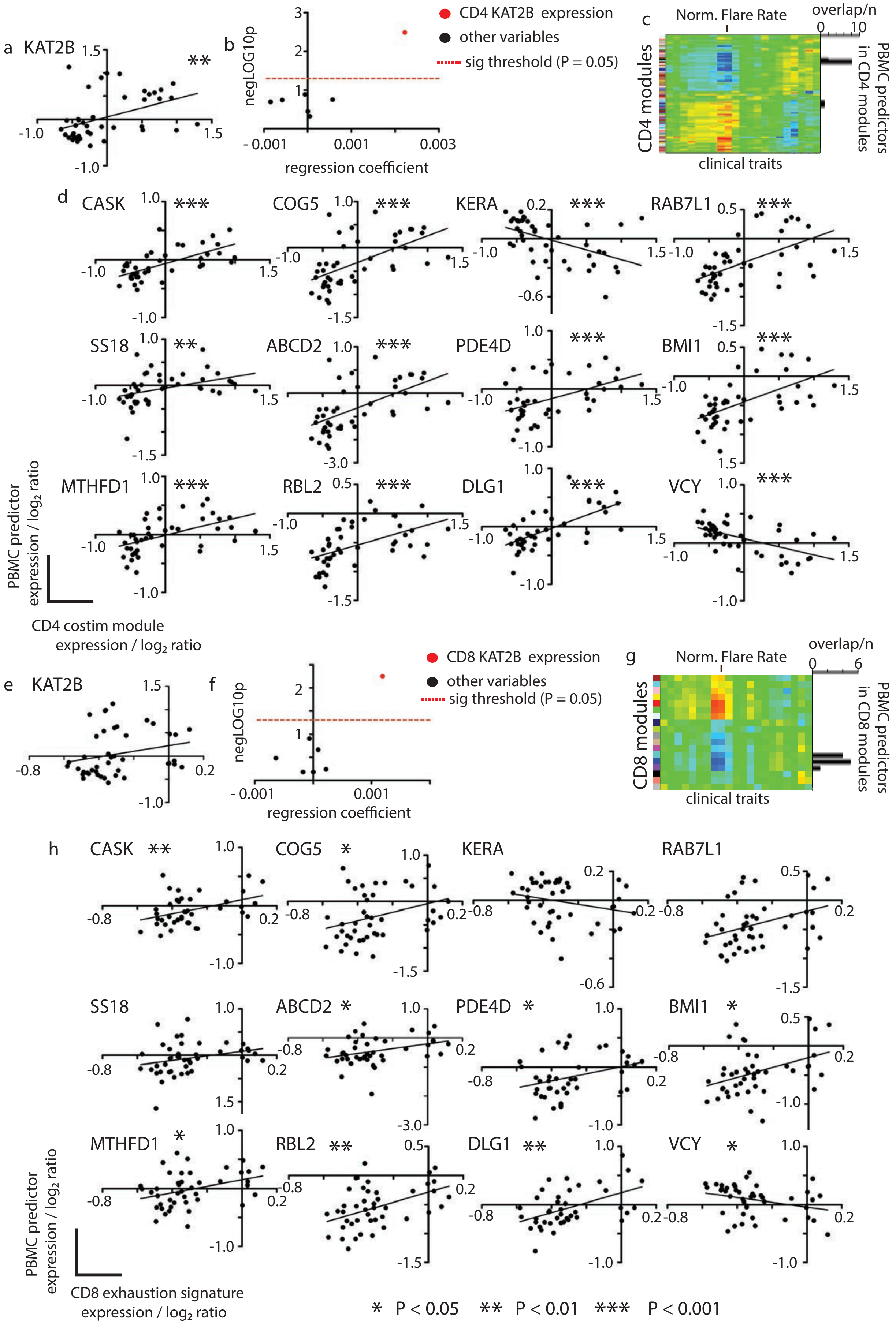
Extended Data Figure 4 | Wind rose plots showing relative GSEA enrichment of immune signatures in autoimmune disease and melanoma. Wind rose plots showing relative enrichment (GSEA FDR *q* value) of distinct immune signatures between subgroups of patients (as defined as in Fig. 2). **a**, **b**, AAV; **c**, **d**, SLE; **e**, **f**, IBD. **a**, **c**, **e**, Enrichment of immune signatures from selected CD8 T-cell phenotypes; **b**, **d**, **f**, enrichment of signatures specifically up-/downregulated by CD8 T-cell subsets derived from the LCMV model of T-cell exhaustion (acute LCMV Armstrong versus chronic LCMV Cl13 (ref. 8)). Detailed information on genes included in each signature is provided in Supplementary Table 6. **g**, **h**, Wind rose plots showing relative enrichment (GSEA FDR *q* value) of distinct immune signatures between CD8 T cells from patients with melanoma, comparing CD8 from tumour-infiltrated lymph node with circulating CD8 T cells[16]. **g**, Enrichment of immune signatures from selected CD8 T-cell phenotypes; **h**, enrichment of signatures specifically up-/downregulated by CD8 T-cell subsets derived from the LCMV model of T-cell exhaustion (acute LCMV Armstrong versus chronic LCMV Cl13 (ref. 8)). Specific enrichment is seen for genes downregulated by exhausted cells but not for all genes upregulated by exhausted cells. **c**, Heat map showing differential expression of selected canonical co-inhibitory receptors (as for Fig. 2c (ref. 12)) in the LCMV exhaustion model, between prognostic subgroups identified in Fig. 2d, g, j and between exhausted CD8 T cells from melanoma-infiltrated lymph node compared with circulating tumour-specific CD8 T cells[16]. Blue, up in exhausted; red, up in non-exhausted; grey, no significant change (FDR $P < 0.05$).


a
unstimulated
92
b
anti-CD3/CD28 only
19
c
anti-CD2/3/28
37
d
+ anti-CD40
17
IL7R
CFSE
e
IL7Rhi (live cells x10^3)
100
10
1
unstim
0
2
4
division #
f
PD-1 (MFI x10^3)
P = 0.0016
8
6
4
2
0
unstim
0
2
4
division #
g
cell death (%)
P < 0.0001
80
60
40
20
0
0
2
4
division #
h
CD4 costimulation
high
low
AAV
SLE
cd4 costimulation module
i
CD8 exhaustion
low
high
AAV
SLE
cd8 exhaustion module
j
gene expression log$_2$ ratio
4
0
-4
4
0
-2
IL7R
PD1
CD2
CD40LG
ICOS
IL7R
PD1
CD2
CD40LG
ICOS
k
gene expression log$_2$ ratio
4
0
-4
2
0
2
IL7R
PD1
CD2
CD40
IL7R
PD1
CD2
CD40
l
CD3/28 + anti-CD40
CD3/28
P = 0.5
IL7R hi cells (x1000)
80
70
40
20
0
unstim
0
2
4
division #
m
CD3/28 + IFNα
CD3/28
P = 0.67
IL7R hi cells (x1000)
80
70
40
20
0
unstim
0
2
4
division #
n
P = 0.2
% live cells
40
20
0
0
2
4
division #
o
P < 0.0001
live cells x 1000
200
150
100
50
0
0
2
4
division #

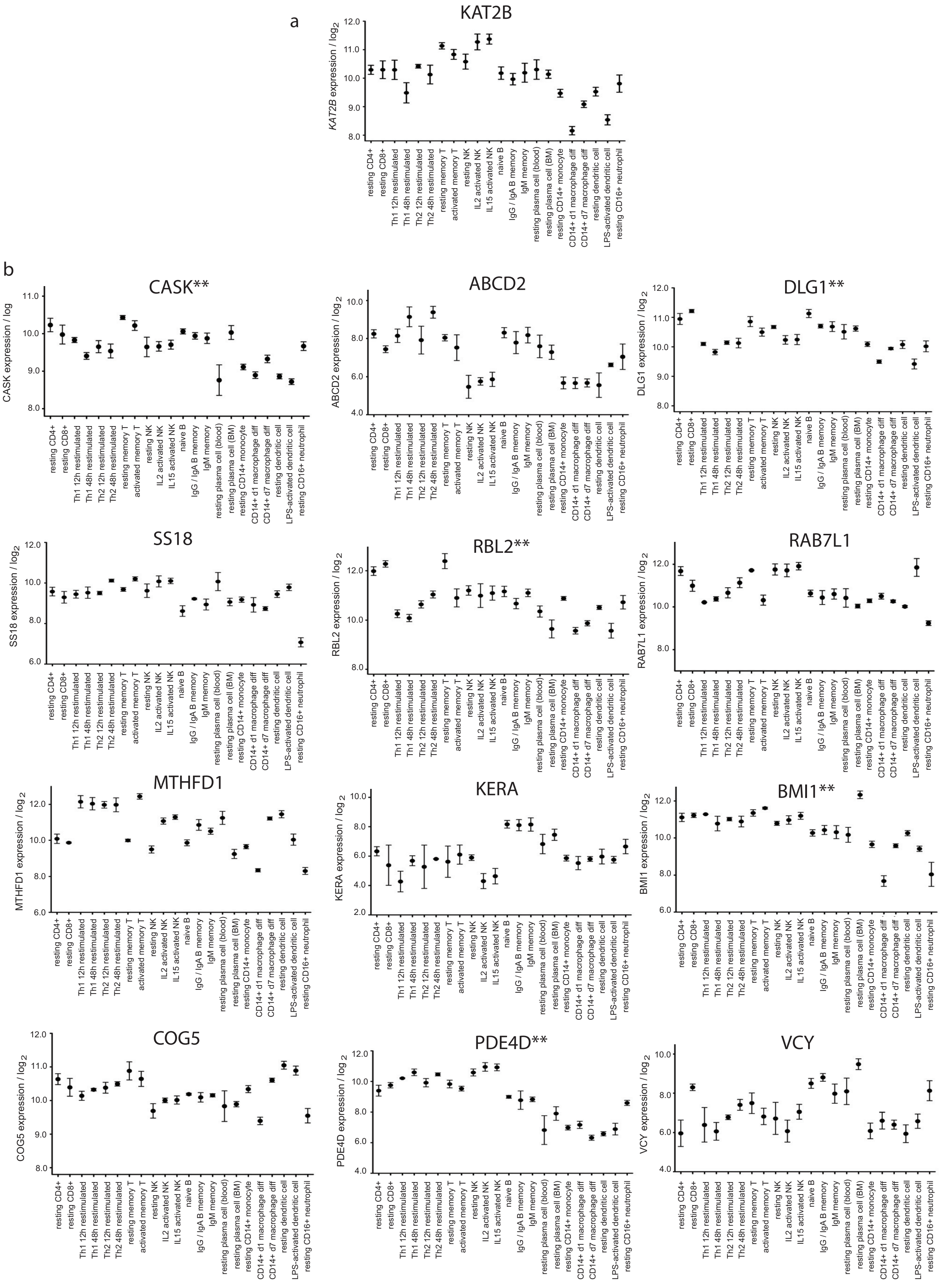
Extended Data Figure 5 | T-cell co-stimulation with CD2, but not type 1 IFN or anti-CD40, prevents development of an exhausted IL-7R^{lo}PD1^{hi} phenotype during prolonged anti-CD3/28 T-cell stimulation. **a**–**d**, Representative scatter plots showing IL-7R expression (*y* axis) by cell division (CFSE dilution, *x* axis) in (**a**) unstimulated cells and following each of three different co-stimulation cultures: **b**, anti-CD3/CD28 alone; **c**, anti-CD2/3/28; **d**, anti-CD40/3/28. IL-7R^{hi}-expressing subset indicated in black gate with the percentage of live cells shown. **e**–**g**, Line and scatter plots showing absolute number of IL-7R^{hi} cells (**e**), PD-1 expression (**f**) and cell death (**g**) (death = AquaFluorescent dye^{+}) during CD8 T-cell differentiation (*x* axis, number of divisions undergone by day 6 of culture measured by CFSE dilution) after anti-CD3/28 (blue) or anti-CD2/3/28 (red) stimulation. *P* = paired *t*-test, *n* = 5 paired samples. **h**, **i**, Hierarchical clustering of 44 patients with AAV (left panels) and 23 patients with SLE (right panels) using 336 genes composing a CD4 T-cell co-stimulation module (black module, Fig. 1) identifies two subgroups of patients (high co-stimulation, red; low co-stimulation, blue) in CD4 T-cell expression data defined by the first major division in the patient dendrogram. **j**, **k**, Scatter plots illustrating selected co-stimulatory and co-inhibitory receptors for the subgroups identified in **h** and **i**. Selected receptors were chosen on the basis of their inclusion in networks derived from the co-stimulation and exhaustion signatures as illustrated in Extended Data Fig. 3a. **l**, **m**, Line and scatter plots showing absolute number of IL-7R^{hi} cells (*y* axis) by number of divisions undergone at day 6 (*x* axis) after polyclonal stimulation with anti-CD3/28 (blue) or anti-CD3/28 plus anti-CD40 (**l**, green) or IFN-α (**m**, green) co-stimulation. **n**, Line and scatter plot showing extent of proliferation occurring (percentage of live cells on day 6 having undergone each of zero to four divisions) after polyclonal stimulation of primary human CD8 T cells with CD3/28 alone (blue) or with additional anti-CD2 co-stimulation (red), confirming no difference in the extent of live cell proliferation between groups. **o**, Absolute live (AquaFluorescent Dye^{-}) cell counts (*y* axis) by the number of divisions undertaken (*x* axis) by day 6 after polyclonal stimulation of primary human CD8 T cells with CD3/28 alone (blue) or with additional anti-CD2 co-stimulation (red), illustrating increased cell survival with CD2 co-stimulation despite equivalent proliferation. *P* values = two-way ANOVA of four paired stimulations.


a
unstimulated
CD3/28
CD2/3/28
BCL2
12
37
12
CFSE
CFSE
CFSE
b
P = 0.025
BCL2 MFI
4000
2000
0
CD2/3/28
CD3/28
c
IL10
P = 0.028
pg/ml
80
40
0
CD3/28
CD2/3/28
TNFα
P = 0.87
pg/ml
15,000
10,000
5000
0
CD3/28
CD2/3/28
IFNψ
P = 0.04
40,000
20,000
0
CD3/28
CD2/3/28
IL1β
P = 0.75
30
20
10
0
CD3/28
CD2/3/28
d
CD3/28
CD2/3/28
IL7R
IL7Rhi
IL7Rlo
IL7R
CFSE
CFSE
e
% live cells
80
60
40
20
0
IL7Rhi
IL7Rlo
IL7Rhi
IL7Rlo
CD2/3/28
CD3/28
f
Tcm
Tn
Tem
Temra
CD62L
CD45RA
g
● anti-CD2/3/28 stimulation
● anti-CD3/28 stimulation
CD8 IL7Rhi / log cells x10
100
10
1
0.1
*
*
Tn
Tcm
Tem
Temra
stimulated Tmem population
h
stimulated: Tcm
% live cells
80
40
0
Tn
Tcm
Tem
Temra
resulting Tmem populations
stimulated: Tn
% live cells
80
40
0
Tn
Tcm
Tem
Temra
resulting Tmem populations
stimulated: Tem
% live cells
80
40
0
Tn
Tcm
Tem
Temra
resulting Tmem populations
stimulated: Temra
% live cells
80
40
0
Tn
Tcm
Tem
Temra
resulting Tmem populations

Extended Data Figure 6 | CD2 co-stimulation results in functionally distinct subpopulations showing enhanced survival after *in vitro* re-stimulation but no preferential expansion of CD8 memory subsets. **a**, Representative flow cytometry density plots of CD8 T cells showing BCL2 expression on day 7 after stimulation with anti-CD3/28 (blue) or anti-CD2/3/28 (red). Figures are the percentage of total CD8 T cells. **b**, Quantification of BCL2 expression in CD8 T cells stimulated as in **a**. $P =$ Mann–Whitney, $n = 5$ paired biological replicates per group. **c**, Scatter plots showing cytokine levels (y axis, picograms per millilitre) measured in supernatants of CD8 T cells on day 7 after *in vitro* stimulation with either anti-CD3/28 (left column, blue) or CD2/3/28 (right column, red). Samples represent paired stimulations of primary CD8 T cells from the same individual using either stimulation protocol ($n = 6$ biological replicates per group). **d**, Scatter plots illustrating populations sorted after polyclonal anti-CD3/28 (left panel) and anti-CD2/3/28 (right panel) stimulation of primary CD8 T cells. **e**, Percentage of live cells (AquaFluorescent dye^{-}) remaining 7 days after re-stimulation of each sorted subpopulation of CD8 cells. Cells were rested for 6 days in complete RPMI1640 medium without IL-2 before being re-stimulated with anti-CD2/3/28 for a further 7 days. $P =$ Mann–Whitney; error bars, mean ± s.e.m. **f**, Representative scatter plot illustrating CD8 T-cell memory populations isolated by flow cytometric sorting and stimulated in **g**, **h**. **g**, Scatter plot showing absolute number of IL-7R$^{\mathrm{hi}}$ cells (y axis) on day 6 after anti-CD3/28 (blue) or anti-CD2/3/28 (red) stimulation of purified CD8 T-cell memory populations (x axis). *$P < 0.05$, Mann–Whitney test ($n = 5$ paired biological replicates per group). **h**, Scatter plots showing percentage CD8 T-cell memory subsets (y axis) resulting from stimulation of purified central memory (Tcm), naive (Tn), effector memory (Tem) and effector memory-RA (Temra) populations with anti-CD3/28 (blue) or anti-CD2/3/28 (red) for 6 days ($n = 4$ paired biological replicates per group).

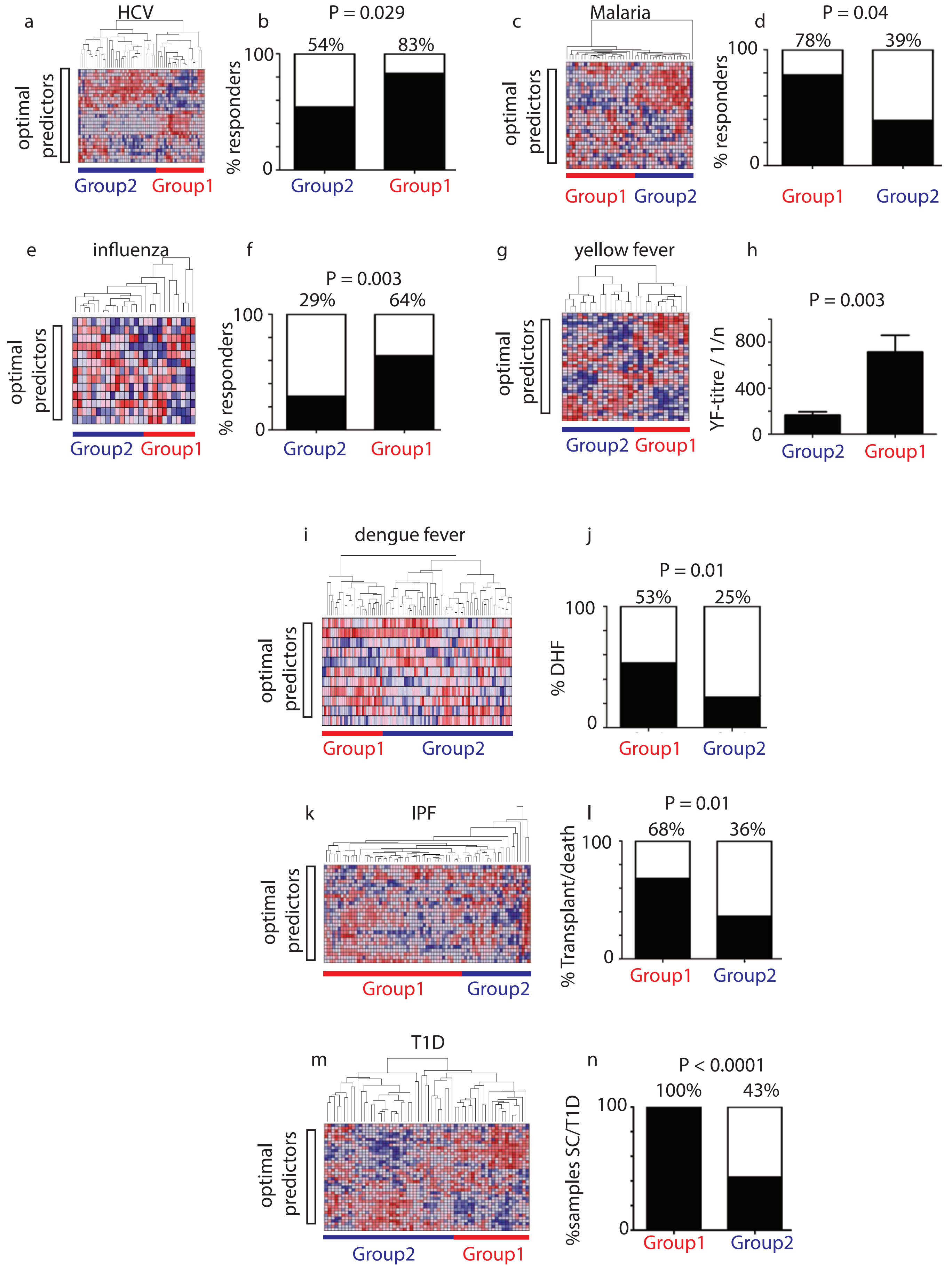

a KAT2B
**
1.5
-1.0
1.5
-1.0
b
negLOG10p
3
2
1
- 0.001
0.001
0.003
regression coefficient
● CD4 KAT2B expression
● other variables
sig threshold (P = 0.05)
c
Norm. Flare Rate
overlap/n
0
10
CD4 modules
clinical traits
PBMC predictors in CD4 modules
d
CASK ***
COG5 ***
KERA ***
RAB7L1 ***
SS18 **
ABCD2 ***
PDE4D ***
BMI1 ***
MTHFD1 ***
RBL2 ***
DLG1 ***
VCY ***
PBMC predictor expression / log2 ratio
CD4 costim module expression / log2 ratio
e KAT2B
1.5
-0.8
0.2
-1.0
f
negLOG10p
2
1
- 0.001
0.001
regression coefficient
● CD8 KAT2B expression
● other variables
sig threshold (P = 0.05)
g
Norm. Flare Rate
overlap/n
0
6
CD8 modules
clinical traits
PBMC predictors in CD8 modules
h
CASK **
COG5 *
KERA
RAB7L1
SS18
ABCD2 *
PDE4D *
BMI1 *
MTHFD1 *
RBL2 **
DLG1 **
VCY *
PBMC predictor expression / log2 ratio
CD8 exhaustion signature expression / log2 ratio
* P < 0.05 ** P < 0.01 *** P < 0.001

Extended Data Figure 7 | Top PBMC surrogate markers reflect expression of CD4 co-stimulation/CD8 exhaustion modules within CD4 and CD8 data respectively. Top PBMC-level predictors ($n = 13$) were selected as indicated in Fig. 4a, and data are shown comparing expression of the optimal predictor (*KAT2B*, **a**, **e**) and of each other top predictor gene (**d**, **h**) in PBMC data compared with expression of the CD4 co-stimulation module eigengene in CD4 data (**a**–**d**) and the CD8 exhaustion signature eigengene in CD8 data (**e**–**h**) for $n = 44$ patients with AAV. Significance of correlation: $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. **b**, **f**, Scatter plots showing the outcome of multiple linear regression models testing the association of *KAT2B* expression in CD4 (**b**) and CD8 (**f**) data (red symbols) directly compared with clinical markers of disease activity (black symbols); *x* axis, magnitude of association (regression coefficient, change in normalized flare rate (flares per number of days follow-up) per unit change in each variable tested); *y* axis, significance of association in multiple regression model; *P*, significance threshold (dashed red line, $P = 0.05$). Clinical variables incorporated were disease activity score (BVAS), C-reactive protein, lymphocyte count, neutrophil count, IgG. **c**, **g**, Heat maps reproduced from Fig. 1a, i, respectively, showing overlap of top PBMC-level predictors with the modular analysis presented for CD4 (**c**) and CD8 (**g**) data in Fig. 1. As expected, surrogate markers showed stronger correlation with the CD4 than the CD8 signature as the algorithm was trained to detect the CD4 co-stimulation module.

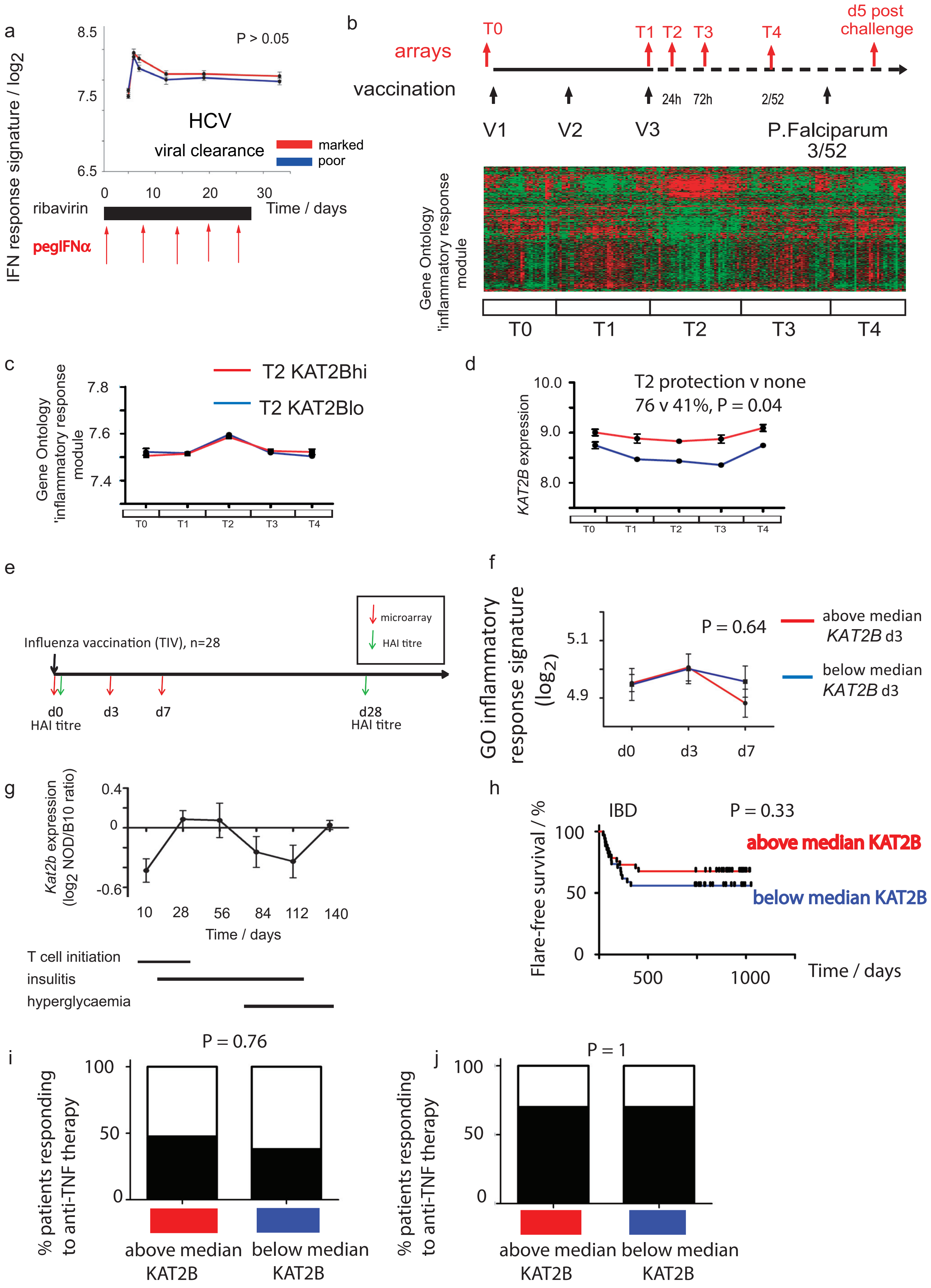


** P < 0.001 correlation with KAT2B expression across all subsets

Extended Data Figure 8 | Immune cell subset expression pattern of top PBMC-level surrogate markers of CD4 co-stimulation/CD8 exhaustion signatures. Dot plots showing expression (median ± s.e.m.) of *KAT2B* (**a**) and for each of 12 other top PBMC-level surrogate predictors of CD4 co-stimulation/CD8 exhaustion signatures (from Fig. 4a) in a range of 22 immune cell subsets. Genes showing significant correlation of expression with *KAT2B* across all cell types are indicated (**$P < 0.001$).


a
HCV
optimal predictors
Group2
Group1
b
P = 0.029
54%
83%
% responders
100
0
Group2
Group1
c
Malaria
optimal predictors
Group1
Group2
d
P = 0.04
78%
39%
% responders
100
0
Group1
Group2
e
influenza
optimal predictors
Group2
Group1
f
P = 0.003
29%
64%
% responders
100
0
Group2
Group1
g
yellow fever
optimal predictors
Group2
Group1
h
P = 0.003
YF-titre / 1/n
800
400
0
Group2
Group1
i
dengue fever
optimal predictors
Group1
Group2
j
P = 0.01
53%
25%
% DHF
100
0
Group1
Group2
k
IPF
optimal predictors
Group1
Group2
l
P = 0.01
68%
36%
% Transplant/death
100
0
Group1
Group2
m
T1D
optimal predictors
Group2
Group1
n
P < 0.0001
100%
43%
%samples SC/T1D
100
0
Group1
Group2

Extended Data Figure 9 | Hierarchical clustering of multiple data sets using 13 top PBMC-level surrogate markers of CD4 co-stimulation/CD8 exhaustion modules identifies subgroups of patients with distinct clinical outcomes. Replication of association between surrogate markers of CD4 co-stimulation/CD8 exhaustion signatures and clinical outcome (as shown in Fig. 4c–k) but using all top 13 PBMC-level surrogates rather than *KAT2B* alone. **a**, **c**, **e**, **g**, **i**, **k**, **m**, Heat maps showing hierarchical clustering of gene expression data of 13 top PBMC-level surrogate predictors of CD4 co-stimulation/CD8 exhaustion signatures (from Fig. 4a) in patients with chronic HCV (**a**), during malaria vaccination (**c**), influenza vaccination (**e**), yellow fever vaccination (**g**), dengue fever infection (**i**), IPF (**k**) and pre-T1D (**m**). Subgroups were defined using a major division of the cluster dendrogram and group 1 allocated on the basis of *KAT2B* expression (highest in group 1). Clinical outcome associated with each subgroup identified is shown in **b** (HCV, percentage of responders to IFN-α/ribavirin therapy), **d** (percentage showing protection versus no protection from malaria vaccine), **f** (percentage response to influenza vaccination), **h** (yellow fever antibody-titre after vaccination), **j** (percentage progression to DHF), **l** (percentage of patients progressing to need for transplantation or death) and **n** (percentage of samples from patients with previous or subsequent progression to islet-cell antibody seroconversion or to a diagnosis of T1D).


a
IFN response signature / log2
8.5
7.5
6.5
P > 0.05
HCV
viral clearance
marked
poor
0
10
20
30
Time / days
ribavirin
pegIFNα
b
arrays
T0
T1
T2
T3
T4
d5 post challenge
vaccination
24h
72h
2/52
V1
V2
V3
P.Falciparum 3/52
Gene Ontology 'inflammatory response module
T0
T1
T2
T3
T4
c
Gene Ontology 'inflammatory response module
7.8
7.6
7.4
T2 KAT2Bhi
T2 KAT2Blo
T0
T1
T2
T3
T4
d
KAT2B expression
10.0
9.0
8.0
T2 protection v none
76 v 41%, P = 0.04
T0
T1
T2
T3
T4
e
microarray
HAI titre
Influenza vaccination (TIV), n=28
d0
HAI titre
d3
d7
d28
HAI titre
f
GO inflammatory response signature (log2)
5.1
4.9
P = 0.64
above median KAT2B d3
below median KAT2B d3
d0
d3
d7
g
Kat2b expression (log2 NOD/B10 ratio)
0.4
0
-0.6
10
28
56
84
112
140
Time / days
T cell initiation
insulitis
hyperglycaemia
h
Flare-free survival / %
IBD
P = 0.33
100
50
0
above median KAT2B
below median KAT2B
500
1000
Time / days
i
% patients responding to anti-TNF therapy
P = 0.76
100
50
0
above median KAT2B
below median KAT2B
j
% patients responding to anti-TNF therapy
P = 1
100
50
0
above median KAT2B
below median KAT2B

Extended Data Figure 10 | Kinetics of *KAT2B* expression during treatment of chronic HCV, malaria and influenza vaccination, during T1D development in the NOD mouse and in PBMC data from patients with IBD and rheumatoid arthritis. a, Expression of a type 1 IFN response signature (average eigenvalue of type 1 IFN response signature plotted for each response group at each time point, A, signature as defined in ref. 4) in a cohort of 54 patients during treatment of chronic HCV infection with pegylated IFN-α and ribavirin (as described in ref. 53 and Fig. 4c), including 28 showing a marked response (red line, HCV titre decrease $>3.5 \log_{10}(\mathrm{IU\,ml}^{-1})$ by day 28) and 26 a poor response (HCV titre decrease $<1.5 \log_{10}(\mathrm{IU\,ml}^{-1})$ by day 28). P = two-way ANOVA. **b**, Schematic representation of the vaccination (black) and transcriptome profiling (red) schedule for the adjuvanted RTS,S malaria vaccine trial[23] (as shown in Fig. 4d). **b**–**d**, Heat map (**b**) and line plots (**c**, **d**) illustrating temporal changes in expression of 404 genes representing the GO 'inflammatory response' module (**c**) or *KAT2B* expression (**d**) at each time-point during vaccination in patients with above- (red) and below- (blue) median *KAT2B* expression throughout the vaccination schedule outlined in **b**. Subgroups defined at T2, immediately after booster vaccination as this equates to the period of most 'active' immune response. Plots are mean ± s.e.m. **e**, Schematic representation of the vaccination (black arrows) and transcriptome profiling (red arrows) schedule for 28 vaccinees receiving the 2008 seasonal influenza vaccination (combined trivalent inactivated influenza vaccine[24] as shown in Fig. 4e) with response assessed at day 28 by haemagglutination inhibition titre (green arrow). **f**, Linear plot illustrating temporal changes in expression of 404 genes representing the GO 'inflammatory response' module at each time-point during vaccination (d0–d7 corresponding to microarray bleed points in **e** for patients showing above- (red) or below- (blue) median expression of *KAT2B* at day 3 after vaccination; *y*, expression $\log_2$; *x*, time-point, days after vaccination; P = two-way ANOVA. **g**, Linear plot showing ratio of *Kat2b* expression in peripheral blood of NOD mice (*y* axis, $n = 37$ mice in total across six time points) before and during the induction and onset of insulitis and the development of overt diabetes (illustrated by black bars below); *x* axis, age (days); *y* axis, *Kat2b* expression $\log_2$ ratio versus B10 controls[29]. **h**, Kaplan–Meier censored survival curve showing flare-free survival (*y* axis) during follow-up (*x* axis) of $n = 58$ patients with IBD stratified by *KAT2B* expression (red, above median; blue, below median). P = log-rank test. **i**, **j**, Box plots showing clinical response (percentage responders) 3 months after treatment with anti-TNF therapy in two independent cohorts (I[54] and J[55]) of patients with rheumatoid arthritis (RA). P = Fisher's exact test. Linear plots show mean ± s.e.m. throughout.

doi:10.1038/nature14614

Mitochondrial reticulum for cellular energy distribution in muscle

Brian Glancy[1]*, Lisa M. Hartnell[2]*, Daniela Malide[1], Zu-Xi Yu[1], Christian A. Combs[1], Patricia S. Connelly[1], Sriram Subramaniam[2] & Robert S. Balaban[1]

Intracellular energy distribution has attracted much interest and has been proposed to occur in skeletal muscle via metabolite-facilitated diffusion[1,2]; however, genetic evidence suggests that facilitated diffusion is not critical for normal function[3–7]. We hypothesized that mitochondrial structure minimizes metabolite diffusion distances in skeletal muscle. Here we demonstrate a mitochondrial reticulum providing a conductive pathway for energy distribution, in the form of the proton-motive force, throughout the mouse skeletal muscle cell. Within this reticulum, we find proteins associated with mitochondrial proton-motive force production preferentially in the cell periphery and proteins that use the proton-motive force for ATP production in the cell interior near contractile and transport ATPases. Furthermore, we show a rapid, coordinated depolarization of the membrane potential component of the proton-motive force throughout the cell in response to spatially controlled uncoupling of the cell interior. We propose that membrane potential conduction via the mitochondrial reticulum is the dominant pathway for skeletal muscle energy distribution.

The mechanism by which the forms of potential energy generated by oxidative phosphorylation are distributed within skeletal muscle cells has been the subject of active research for many years. Two major facilitated diffusion mechanisms have been proposed. The creatine kinase shuttle system proposes that creatine phosphate and creatine are cytosolic facilitated diffusion partners with ATP and ADP[1], while the second model proposes that the oxy-deoxy myoglobin shuttle aids diffusion of oxygen from capillaries to mitochondria[2]. However, in the absence of myoglobin[3], creatine kinase[4], or creatine[5], mice survive with near-normal skeletal muscle performance and only modest adaptations[5,7], suggesting that these facilitated diffusion pathways are not critical for normal muscle function. Thus, a re-examination of the pathways for distributing potential energy within skeletal muscle is warranted.

We selected high-resolution, three-dimensional (3D, 15-nm isotropic voxels) focused ion beam scanning electron microscopy (FIB-SEM)[8,9] to better define the diffusion pathways in mitochondria-rich muscle fibres. Previous investigations of muscle 3D structure[10–15] have yielded conflicting results, probably due to the size of the serial sections (60–90 nm), a lack of true 3D information[16], and/or differences between muscle types studied[12].

The contrast in the FIB-SEM image stack (Supplementary Video 1, raw data available at http://labalaban.nhlbi.nih.gov/files/SuppDataset.tif, see also Supplementary Information) was sufficient to enable automated segmentation of mitochondria, myofibrils, blood vessels, red blood cells and nuclei (Fig. 1a, b) and revealed four different mitochondrial morphologies (Supplementary Video 1): paravascular mitochondria (PVM) (Fig. 1c), I-band mitochondria (IBM) (Fig. 1c–e), fibre parallel mitochondria (FPM, Fig. 1d, e), and cross fibre connection mitochondria (CFCM, Fig. 1d). Consistent with our previous *in vivo* microscopy studies[7,17,18], the PVM pool, often associated with a nucleus, was evident in mitochondria-rich fibres when imaged by FIB-SEM. Tracing the path of the PVM often revealed multiple or single direct connections (that is, a continuous outer membrane) with IBM[19] (Fig. 2b). A PVM could be coupled to one or both IBM on opposing sides of the *z*-line (Supplementary Video 2). CFCM were less frequent and also seen as pairs of tubules proceeding transversely on both sides of a *z*-line. However, CFCM were oriented perpendicularly to IBM

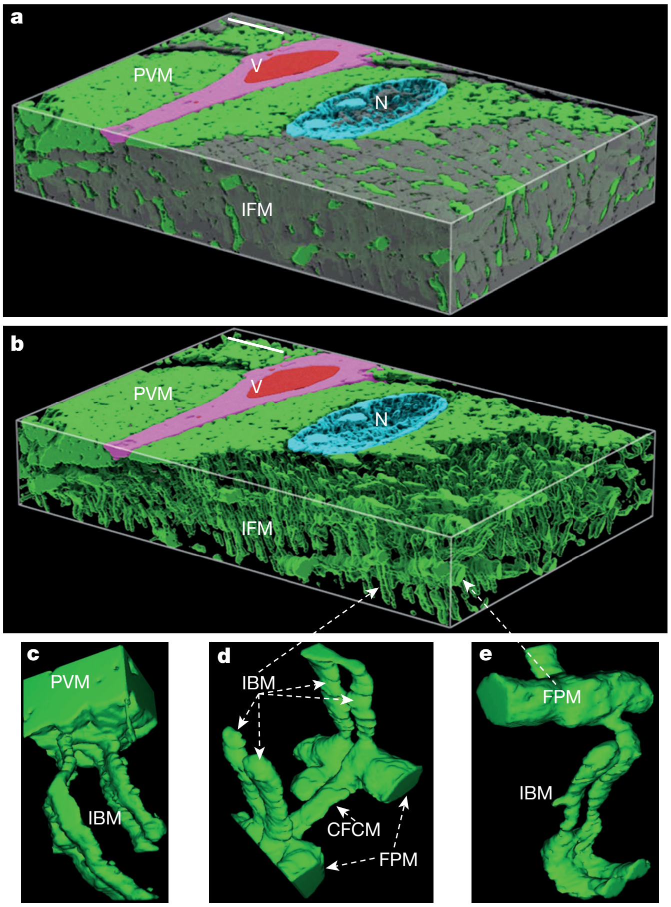


Figure 1 | Muscle mitochondria form highly connected networks. a, 3D surface rendering of 25.53 × 24.06 × 4.23 μm FIB-SEM volume segmented to show spatial relationships between mitochondria (green) and other structures (nucleus (N), cyan; capillary (V), magenta; red blood cell, red; myofibrils, grey). **b**, Removing myofibrils highlights different morphologies within intrafibrillar mitochondrial (IFM) network. **c–e**, Zooming in reveals projections from paravascular mitochondria (PVM) into I-band mitochondria (IBM) (**c**), and numerous interactions between IBM and cross-fibre connection mitochondria (CFCM) (**d**) and fibre parallel mitochondria (FPM) (**d**, **e**). Scale bars, 3 μm. Representative of eight separate volumes analysed from four animals.

[1] National Heart Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [2] National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.

*These authors contributed equally to this work.

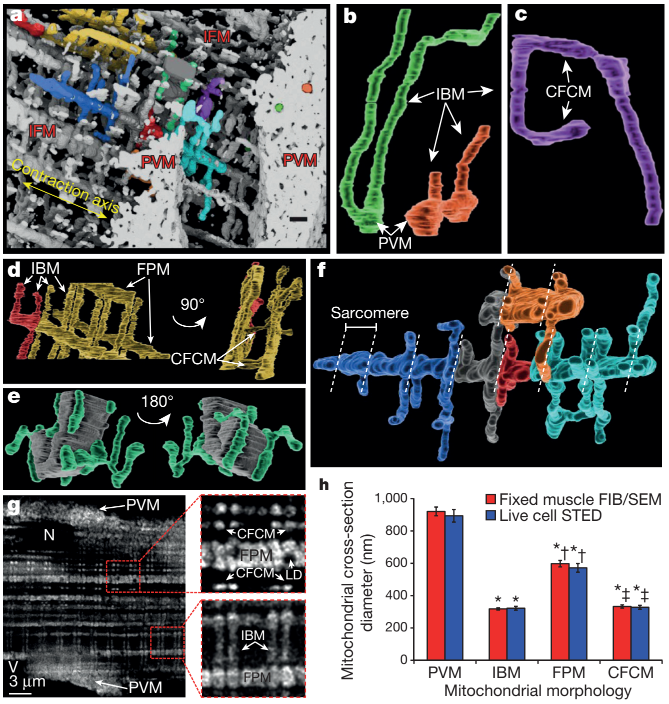


Figure 2 | Muscle mitochondrial morphologies. **a**, 3D rendering of mitochondria from FIB-SEM volume. Non-white colours indicate individual mitochondria. Scale bar, 750 nm. **b**, PVM projecting into IBM. **c**, Mitochondrion showing IBM and CFCM. **d**, Repeating connections between IBM and FPM. **e**, Myofibril (grey) and associated mitochondria (green). **f**, FPM with IBM and CFCM branches. Dotted lines indicate *z*-lines. **g**, Deconvolved super-resolution image of TMRM-loaded myocyte. LD, lipid droplet. Insets are magnified 3× from original. **h**, Mean mitochondrial diameters. Error bars indicate standard error. Quantification from raw images. Significantly different from PVM (*), IBM (†), FPM (‡) (ANOVA, $P < 0.05$). Images represent data from: FIB-SEM, 8 fibres, 4 mice; STED, 13 fibres, 3 mice.

(Fig. 2c, d). Sarcomeres were observed to have associations with multiple mitochondrial structures (Fig. 2e). This analysis demonstrates that PVM are directly coupled to most IBM in the muscle, providing a conductive pathway for potential energy transfer from the periphery to deep inside the muscle. Skulachev proposed over 40 years ago that the electrical component of the proton-motive force, the membrane potential ($\Delta\Psi$), could be used by the cell as a transportable form of potential energy[20] and later demonstrated that mitochondria could be electrically coupled in cultured cells[21].

PVM were highly irregular (Supplementary Video 1), mostly ellipsoidal structures. Their maximal cross-sectional diameter ranged from 300 to 1,600 nm (Extended Data Table 1). PVM that were contiguous with IBM (21.2 ± 3.2% of 1,152 PVM analysed, 4 animals) were located primarily at the PVM–myofibril interface (Extended Data Fig. 1a). Although a small fraction of PVM was directly coupled to IBM, the entire PVM pool (99.9 ± 0.1%) was linked via electron-dense contact sites (EDCS) between membranes of adjacent PVM (Extended Data Fig. 1b). These apparently dynamic EDCS[22,23] may represent sites of fission or fusion[24] as well as simply reflect tight packing of the mitochondria; however, the functional significance of these junctions is poorly defined. If EDCS are electrically conductive, these sites would couple essentially all PVM directly to IBM running deep into the muscle, creating a functional syncytium with regard to the proton-motive force ($\Delta\Psi$ + pH gradient across the inner mitochondrial membrane) to generate ATP.

IBM were regular cylinders with a maximal cross-sectional diameter range of <100–678 nm (Extended Data Table 1) and, often, lengths >20 µm (Fig. 2), suggesting that they extend deep into the tissue. FPM (Fig. 2d, f) displayed maximal cross-sectional diameters ranging from 362 to 1,495 nm (Extended Data Table 1). 99.7 ± 0.3% of FPM ($n = 776$, 4 FIB-SEM data sets, 4 animals) were associated with adjacent FPM through EDCS (Extended Data Fig. 1d). Interconnection of different intra-fibrillar mitochondrial (IFM) morphologies was extensive. Indeed, 81.7 ± 3.6% of FPM were directly connected (continuous outer membrane) to an IBM (Extended Data Fig. 1c, $n = 305$ mitochondria from 4 FIB-SEM data sets from 4 animals). Mitochondria lacking either direct or EDCS connections were rare (~1–2 per data set) and may reflect either damaged mitochondria removed from the network[11] or newly formed mitochondria yet to establish a network connection[10,11]. For a better appreciation of the mitochondrial reticulum throughout the entire cell, we collected 3D, multiphoton microscopy (MPM) images of mitochondria-rich fibres *in situ*. Although the spatial resolution was ~20-fold lower than with FIB-SEM, these MPM images (Extended Data Fig. 2 and Supplementary Videos 3, 4 and 5) suggest that the morphology observed by FIB-SEM at the periphery of the fibre persists throughout the entire cell. Additionally, super-resolution microscopy of mitochondria using $\Delta\Psi$-dependent tetramethylrhodamine (TMRM) in live, isolated muscle fibres provided further validation that the mitochondrial morphologies observed with FIB-SEM are representative of the *in vivo* condition (Fig. 2g, h, Extended Data Fig. 3 and Supplementary Video 6).

On the basis of the mitochondrial structures, we hypothesized that PVM are primarily involved in generation of the proton-motive force near the capillaries, while IBM use the proton-motive force to produce ATP near the ATPase activity. Along this line of thought, the distribution of the enzymes associated with production and utilization of the proton-motive force may be linked to the function of PVM and IBM pools, thereby optimizing the regional inner membrane concentration of enzymes to its functional requirements. Again, the PVM generation of $\Delta\Psi$ is delivered to the IBM system via the mitochondria reticulum. To test this hypothesis, we compared the cellular distribution of a proton-motive-force-generating element, complex IV, with the enzyme complex responsible for phosphorylation of ADP to ATP using the proton-motive force, complex V. Figure 3a–c shows that complex IV (green) was higher in the periphery of the cell, consistent with the PVM pool, while complex V (red) was more concentrated in the intra-fibrillar region (Fig. 3 and Supplementary Video 7). Notably, complex IV and complex V were found throughout the cell (Supplementary Video 8 and Extended Data Fig. 4), just at different proportions in the different mitochondrial pools. Antibody specificity was validated by the appearance of single bands corresponding to complexes IV and V in western blots of mitochondrial proteins separated by clear native PAGE (Extended Data Fig. 5).

To test our hypothesis that the mitochondrial reticulum is coupled via a contiguous matrix and/or EDCS, we performed localized mitochondrial uncoupling experiments and observed whether the associated drop in $\Delta\Psi$ propagated throughout the mitochondrial pools in isolated muscle cells. In this study, we simulated an increase in intra-fibrillar complex V activity with a photo-activated uncoupler[25] and observed the cellular topology of $\Delta\Psi$ via TMRM distribution. Upon a mild depolarization of a small interior cell region, a rapid (<200 ms) and near-uniform drop in $\Delta\Psi$ occurred across the cell, including a homogenous decrease in the PVM, seen as a redistribution of TMRM from mitochondria to cytosol and nuclei (Fig. 4 and Supplementary Video 9). Note that capillary grooves in the sarcolemma and adjacent PVM[7] are still present even after vessel removal by the fibre isolation process (Extended Data Fig. 6). Figure 4k shows the mean post/pre change in TMRM intensity for mitochondrial and cytosolic plus nuclear pixels in each cell region. This analysis resulted in isolated signals from the cytosol, nucleus and PVM pools. However, all of the intra-fibrillar mitochondria were grouped together in this analysis. While linear intensity analysis (Extended Data Fig. 7a–d) could be used to assess IBM in some cases, any inclusion of FPM in the analysis window would be confounding, thereby limiting the image regions available for analysis. We were able to isolate the entire IBM pool based on their

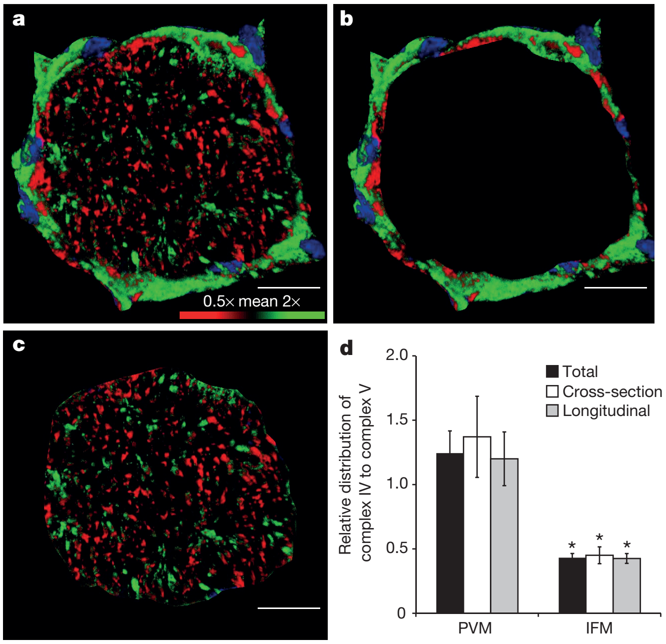


Figure 3 | Capacity for membrane potential conduction. **a**, 3D rendering of muscle fibre immunostained for both complex IV and complex V. Confocal image coloured according to complex IV/complex V ratio. Relatively higher complex IV, green pixels; relatively higher complex V, red pixels. Nuclei, blue. **b**, **c**, Separation of PVM around fibre periphery (**b**) and IFM (**c**) within fibre as used for calculations. **d**, PVM have relatively greater capacity for membrane potential generation while IFM have greater capacity for membrane potential utilization. Images are representative of data from 12 fibres, 5 mice. Error bars indicate standard error. Asterisk indicates significantly different from PVM (paired *t*-test, $P < 0.05$). Scale bars, 15 µm.

regular ~2 µm spacing (or 0.5 µm^{-1} spatial frequency) using a horizontal fast Fourier transform (FFT). A decrease (61 ± 3% ($n = 9$)) in the FFT-detected IBM pool occurred with a more even TMRM distribution between the IBM and cytosol (Extended Data Fig. 7e, f). Conversely, ultraviolet (UV) exposure in control cells resulted in a small, uniform drop in TMRM in both the cytosol and mitochondria in the stimulated region, consistent with slight TMRM photobleaching, and little effect in other regions of the cell, including the IBM pool (Extended Data Figs 7f and 8 and Supplementary Video 10). Restriction of UV light to designated regions of interest (ROIs) was confirmed in fluorescent test slides (Extended Data Fig. 9).

These data are consistent with a tight, rapid electrical coupling mechanism between mitochondria, and that the PVM pool is a functional syncytium with regard to $\Delta\Psi$. The rapid, homogeneous response to the modest uncoupling, as shown in Fig. 4 and Supplementary Video 9, is indicative of a highly coupled, conductive network, although the mechanisms for this are yet to be fully resolved. We suggest, based on our structural data, that the contiguous matrix elements between mitochondrial pools are one of the obvious conductive elements of the network. We speculate that the EDCS are also conductive elements specifically within the PVM and FPM pools. Mitochondrial fission and fusion[26] may also have a role in reticulum development and modulation. It is important to note that the putative conduction through the EDCS could also be dynamically modulated to affect mitochondrial electrical coupling.

Energy requirements in skeletal muscle can nearly instantaneously increase by more than 100-fold during intense contraction[27]. Thus, using electrical conduction as opposed to oxygen or ATP diffusion is a more effective way to quickly and uniformly distribute energy throughout these cells and may explain why disruption of the known ATP and oxygen facilitated diffusion systems results in only modest phenotype changes[3,6,7]. If electrical conduction via the mitochondrial

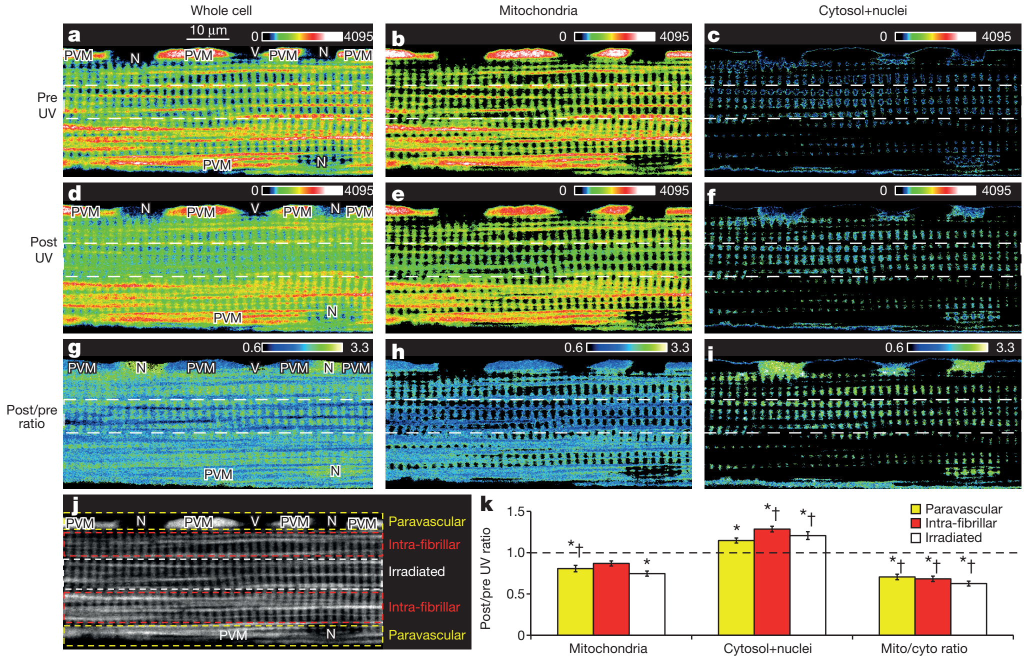


Figure 4 | Mitochondrial membrane potential conduction. Regional uncoupling of fibre loaded with TMRM and MitoPhotoDNP. White dotted lines indicate region of MitoPhotoDNP UV activation. **a**–**c**, Pre-UV whole-cell (**a**), mitochondria (**b**) and cytosol plus nuclei (**c**) images. **d**–**f**, Post-UV whole-cell (**d**), mitochondria (**e**) and cytosol plus nuclei (**f**) images. **g**–**i**, Post/pre-UV ratios for whole-cell (**g**), mitochondria (**h**) and cytosol plus nuclei (**i**). **j**, Cell regions. **k**, Near-uniform redistribution of TMRM from mitochondria to cytosol plus nuclei consistent with homogenous depolarization of all cell regions. Mean ± standard error from 11 experiments, 4 animals. Significantly different (ANOVA, $P < 0.05$) values are indicated for post/pre ratio of 1.0 (*), control (†) (Extended Data Fig. 8). No statistical differences between cell regions.

reticulum is the major mechanism for potential energy distribution within the cell, this raises the question of why such a large investment in creatine kinase, creatine and myoglobin has been retained by the muscle. In the knockouts of these systems, small decrements, ~10%, in peak performance have been noted. Although these are often discounted as small effects, a 10% advantage in peak muscle performance would probably provide a significant evolutionary advantage, justifying the retention or development of these two complementary systems. As a reference, athletic performance is often measured in fractions of a per cent; evolutionary advantages might have a similar sensitivity. We propose that this conductive pathway is the major mechanism for energy distribution in skeletal muscle cells under normal conditions, and the metabolite facilitated diffusion pathways only become significant as one approaches maximum performance levels.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 August 2014; accepted 18 May 2015.

1. Bessman, S. P. & Geiger, P. J. Transport of energy in muscle: The phosphorylcreatine shuttle. *Science* **211,** 448–452 (1981).
2. Wittenberg, J. B. Myoglobin-facilitated oxygen diffusion: Role of myoglobin in oxygen entry into muscle. *Physiol. Rev.* **50,** 559–632 (1970).
3. Garry, D. J. *et al.* Mice without myoglobin. *Nature* **395,** 905–908 (1998).
4. van Deursen, J. *et al.* Skeletal muscles of mice deficient in muscle creatine kinase lack burst activity. *Cell* **74,** 621–631 (1993).
5. Lygate, C. A. *et al.* Living without creatine: unchanged exercise capacity and response to chronic myocardial infarction in creatine-deficient mice. *Circ. Res.* **112,** 945–955 (2013).
6. Kernec, F., Unlu, M., Labeikovsky, W., Minden, J. S. & Koretsky, A. P. Changes in the mitochondrial proteome from mouse hearts deficient in creatine kinase. *Physiol. Genomics* **6,** 117–128 (2001).
7. Glancy, B. *et al. In vivo* microscopy reveals extensive embedding of capillaries within the sarcolemma of skeletal muscle fibers. *Microcirculation* **21,** 131–147 (2014).
8. Narayan, K. *et al.* Multi-resolution correlative focused ion beam scanning electron microscopy: Applications to cell biology. *J. Struct. Biol.* **185,** 278–284 (2014).
9. Dahl, R. *et al.* Three dimensional reconstruction of the human skeletal muscle mitochondrial network as a tool to assess mitochondrial content and structural organization. *Acta Physiol.* **213,** 145–155 (2015).
10. Bakeeva, L. E., Chentsov, Y. S. & Skulachev, V. P. Ontogenesis of mitochondrial reticulum in rat diaphragm muscle. *Eur. J. Cell Biol.* **25,** 175–181 (1981).
11. Bakeeva, L. E., Chentsov Yu, S. & Skulachev, V. P. Mitochondrial framework (reticulum mitochondriale) in rat diaphragm muscle. *Biochim. Biophys. Acta* **501,** 349–369 (1978).
12. Kayar, S. R., Hoppeler, H., Mermod, L. & Weibel, E. R. Mitochondrial size and shape in equine skeletal muscle: a three-dimensional reconstruction study. *Anat. Rec.* **222,** 333–339 (1988).
13. Kirkwood, S. P., Munn, E. A. & Brooks, G. A. Mitochondrial reticulum in limb skeletal muscle. *Am. J. Physiol.* **251,** C395–C402 (1986).
14. Kirkwood, S. P., Packer, L. & Brooks, G. A. Effects of endurance training on a mitochondrial reticulum in limb skeletal muscle. *Arch. Biochem. Biophys.* **255,** 80–88 (1987).
15. Ogata, T. & Yamasaki, Y. Scanning electron-microscopic studies on the three-dimensional structure of mitochondria in the mammalian red, white and intermediate muscle fibers. *Cell Tissue Res.* **241,** 251–256 (1985).
16. Denk, W. & Horstmann, H. Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure. *PLoS Biol.* **2,** e329 (2004).
17. Bakalar, M. *et al.* Three-dimensional motion tracking for high-resolution optical microscopy, *in vivo*. *J. Microsc.* **246,** 237–247 (2012).
18. Rothstein, E. C., Carroll, S., Combs, C. A., Jobsis, P. D. & Balaban, R. S. Skeletal muscle NAD(P)H two-photon fluorescence microscopy *in vivo*: topology and optical inner filters. *Biophys. J.* **88,** 2165–2176 (2005).
19. Bubenzer, H. J. Die Dunnen Und Die Dicken Muskelfasern Des Zwerchfells Der Ratte. *Z. Zellforsch. Mikrosk. Anat.* **69,** 520–550 (1966).
20. Skulachev, V. P. Energy transformation in the respiratory chain. *Curr. Top. Bioenerg.* **4,** 127–190 (1971).
21. Amchenkova, A. A., Bakeeva, L. E., Chentsov, Y. S., Skulachev, V. P. & Zorov, D. B. Coupling membranes as energy-transmitting cables. I. Filamentous mitochondria in fibroblasts and mitochondrial clusters in cardiomyocytes. *J. Cell Biol.* **107,** 481–495 (1988).
22. Picard, M. *et al.* Acute exercise remodels mitochondrial membrane interactions in mouse skeletal muscle. *J. Appl. Physiol.* **115,** 1562–1571 (2013).
23. Saito, A., Smigel, M. & Fleischer, S. Membrane junctions in the intermembrane space of mitochondria from mammalian tissues. *J. Cell Biol.* **60,** 653–663 (1974).
24. Eisner, V., Lenaers, G. & Hajnoczky, G. Mitochondrial fusion is frequent in skeletal muscle and supports excitation-contraction coupling. *J. Cell Biol.* **205,** 179–195 (2014).
25. Chalmers, S. *et al.* Selective uncoupling of individual mitochondria within a cell using a mitochondria-targeted photoactivated protonophore. *J. Am. Chem. Soc.* **134,** 758–761 (2012).
26. Fan, X., Hussien, R. & Brooks, G. A. H_2O_2-induced mitochondrial fragmentation in C_2C_{12} myocytes. *Free Radic. Biol. Med.* **49,** 1646–1654 (2010).
27. Weibel, E. R. & Hoppeler, H. Exercise-induced maximal metabolic rate scales with muscle aerobic capacity. *J. Exp. Biol.* **208,** 1635–1644 (2005).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to thank E. Tyler for his assistance with image rendering, J. Taraska for providing experimental advice, T. Karpova and the NCI Core Fluorescence Imaging Facility for access to a microscope with a 355-nm laser, and P. Kellman for assistance with image analysis. We would also like to thank S. Caldwell and R. Hartley for providing us with MitoPhotoDNP. This study was supported by intramural funding of the Division of Intramural Research, National Heart, Lung, and Blood Institute and the Center for Cancer Research, National Cancer Institute.

Author Contributions R.S.B., S.S., B.G. and L.M.H. designed and L.M.H performed the FIB-SEM experiments. R.S.B. and B.G. designed and B.G. performed the MPM experiments. R.S.B., B.G., Z.-X.Y. and D.M. designed and B.G., Z.-X.Y. and D.M. performed the dual immunolabelling experiments. P.S.C. performed the TEM imaging. R.S.B. and B.G. performed the segmentations and analysed the data. R.S.B., C.A.C. and B.G. designed and performed the isolated muscle fibre experiments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.S.B. (Robert.Balaban@nih.gov).

METHODS

Mice. Male C57BL/6 mice, 2–4 months old, were purchased from Taconic Farms (Germantown, NY). All mice were fed ad libitum and kept on a 12 h light, 12 h dark cycle at 20–26 °C. The experiments were not randomized; the investigators were not blinded to allocation during experiments and outcome assessment.

Animal preparation. All procedures were approved by the National Heart, Lung, and Blood Institute Animal Care and Use Committee and performed in accordance with the guidelines described in the Animal Care and Welfare Act (7 USC 2144). Mice were anaesthetized and the tibialis anterior (TA) muscle prepared by removing the skin and outer layers of fascia as described previously[28].

Focused ion beam scanning electron microscopy (FIB-SEM). Mouse hindlimbs were stabilized in the lengthened position and submerged *in vivo* in 5% glutaraldehyde in 0.12 M sodium cacodylate, pH 7.35 (fixative) for 15 min. After initial fixation, the TA was excised, placed into fresh, cold fixative, and cut into small strips. Fixed muscles were post-fixed with 1% osmium tetroxide, stained with 1% uranyl acetate, dehydrated with an ethanol series and propylene oxide, and embedded in EMbed-812 (Electron Microscopy Sciences, Hatfield, PA).

Nine data sets of mouse tibialis anterior muscle from four animals were collected on mitochondria-rich skeletal muscle fibres. One data set failed due to technical difficulties during data collection. A Zeiss Nvision 40 microscope was used for FIB-SEM imaging. The areas of muscle were chosen for FIB-SEM data collection following survey of 0.5–1-μm-thick sections of resin-embedded muscle tissue; sections were created using an Ultracut S microtome from Leica Microsystems. The sections were stained with Azure II/Toluidine blue that imparts colour to common morphological elements such as nuclei, plasma membranes and mitochondria. Once stained, the orientation and morphology were assessed using a light microscope. Muscle fibres were selected based on the presence of large accumulations of mitochondria in the paravascular region adjacent to a capillary embedded in the muscle cell[7]. Muscle fibres in the correct orientation with good morphology were chosen for FIB-SEM data collection and digital images were taken at 10×, 40× and 60× magnification. These images were used as maps to pinpoint the previously selected areas for data collection in the FIB-SEM.

SEM images were collected with 1.5 kV landing energy using the ESB (backscattered electron) detector. Parameters were chosen to provide 8-bit images with pixel sizes $x = 5$ nm, pixel size $y = 5$ nm, and z thickness = 15 nm. The FIB aperture sizes used for data collection were 300 pA and 700 pA. Automated data collection was performed using Atlas 3D software, which is integrated into the Zeiss platform, manufactured by Fibics[29].

Focused ion beam scanning electron microscopy image processing and analysis. Images (tifs) of each data set were aligned to each other using a cross-correlation algorithm[30], then binned 3 × 3 in x and y directions to create isotropic volumes with edge dimensions of 15 nm. These volumes were initially analysed using IMOD open source software package[31]. Whole image segmentation was performed by first applying a 3D, two pixel median filter on a FIB-SEM data set and then manually adjusting the threshold to create binary images primarily including only mitochondria, myofibrils, nuclei, blood vessels, red blood cells, and/or z-lines. Thresholding often resulted in inclusion of two or more types of cellular structures. Individual structures were then segmented by subtracting images that had been differentially thresholded and/or using the Remove Outliers tool in ImageJ (National Institutes of Health, Bethesda, MD).

Using the TRAKEM2 plugin in ImageJ, the mitochondrial volumes were manually traced throughout the 3D volumes by two different observers. Most of the IBM structures were limited by the field of view of the 3D volumes, thus, the long axis extent of these structures through the cell were potentially grossly underestimated. For studies designed to estimate the number of PVM or FPM that have direct coupling to the IBM (that is, a continuous outer membrane), tracking was initiated at the middle of the 3D stack to minimize the effect of the field of view on detecting complete mitochondrial volumes. All PVM or FPM were then tracked through the 3D volume. Those mitochondria with a continuous projection into an IBM were labelled with a green spot. Those without a projection to an IBM were labelled with a red spot. PVM or FPM that continued out of the field of view without an IBM projection were labelled with a blue spot.

To determine the potential network of coupling between the mitochondria via EDCS, the same mid-volume planes were used. Initiating in the centre of a PVM pool or in the intra-fibrillar space for FPM analysis, every mitochondrion that was 'connected' via an EDCS was marked with a magenta marker. The PVM with IBM projections were also labelled with a green marker. Thus, PVM that project into IBM and also have EDCS were labelled with both green and magenta dots. The EDCS coupled most PVM and most FPM even without looking through the entire 3D volume. However, when using the full 3D volume, nearly all of the mitochondria were coupled via the EDCS structures. Mitochondria not coupled via an EDCS were labelled in orange in these studies; uncoupled mitochondria were rarely observed.

Transmission electron microscopy. Selected specimens for FIB-SEM imaging were sectioned to produce ∼100–150-nm-sized thin sections that were stained with uranyl acetate and lead citrate before imaging on a JEM 1400 electron microscope (JEOL USA, Peabody, MA) with an AMT XR-111 digital camera (Advanced Microscopy Techniques Corporation, Woburn, MA).

Multi-photon image acquisition. Mouse TA muscles were imaged *in situ* after cutting the hindlimb just above the knee, leaving the lower hindlimb and the origin and insertion of the TA intact. The lower hindlimb was placed in a deformable cast in a Petri dish with the TA in the lengthened position and immersed in phosphate-buffered saline. Images were acquired using a Leica TCS SP5 II upright, resonant scanning, multi-photon microscope with a Nikon 25X, 1.1 NA water immersion objective. The exposed muscle was imaged with a pulsed Ti:sapphire laser tuned to 720 nm (Spectra Physics, Irvine, CA, USA), and all emitted light was collected by two hybrid detectors (HyDs, Leica Microsystems) separated by a 545 nm dichroic mirror. 3D image stacks of endogenous NAD(P)H were collected as 1,024 × 1,024 pixel, 12-bit images with 100 nm x and y pixel sizes and 100 nm z-steps between images. Image stacks were deconvolved using the measured point spread function and the 3D Parallel Iterative Deconvolution plugin in ImageJ for presentation purposes only.

Dual muscle section immunostaining. Mouse TA muscles were frozen in the lengthened position and embedded in optimal cutting temperature compound (OCT). Cryosections were cut and air dried for 5 min before fixation in 10% buffered formalin for 7 min at room temperature for confocal imaging. Samples were washed three times in PBS for 5 min each and permeabilized with 0.01% Triton-X 100 in PBS for 5 min. Sections were blocked with 10% goat serum for 20 min before incubation with primary antibodies. Distribution of Complex V was assessed using anti-ATPB rabbit polyclonal IgG antibody from Abcam (ab128743; Cambridge, MA) and complex IV distribution was assessed with anti-complex IV subunit I mouse monoclonal IgG2a from Life Technologies (1D6E1A8; Grand Island, NY). Primary antibodies were incubated overnight at 4 °C at 1:75 (complex V) and 1:40 (complex IV) dilutions. After three 5 min washes in PBS, samples were incubated with secondary antibodies for 1 h at room temperature. Secondary antibodies used were TRITC goat, anti-rabbit IgG (1:100) and FITC goat, anti-mouse IgG (1:100) from Jackson ImmunoResearch Laboratories (West Grove, PA) for confocal imaging. To reduce non-specific binding of the anti-mouse secondary antibody, the MOM Fluorescein Kit from Vector Laboratories (Burlingame, CA) was used. Samples were then washed with PBS and mounted with VectaShield and DAPI (Vector) for confocal imaging. Slides were prepared for sections with each primary antibody separately, both antibodies together, and with primary antibodies omitted (negative control) from 6 TA muscles from 5 mice. No statistical methods were used to predetermine sample size—with the limited information available, we selected the sample size based on prior experience with immunostaining procedures in muscle.

Immunostained muscle image acquisition. For confocal imaging, stacks of high resolution, 12 bit images (100 nm x and y pixel size) were collected sequentially throughout the depth of skeletal muscle tissue sections at 0.2 μm depth intervals using a 63× 1.4 NA oil immersion objective on a Leica SP5 confocal system (Leica Microsystems, Mannheim, Germany), or with a 63× 1.4 NA oil immersion objective on a Zeiss LSM780 confocal microscope (Carl Zeiss MicroImaging, Jena, Germany). FITC-labelled samples were imaged with a 488-nm excitation laser and a 505–550 nm acquisition window and TRITC-labelled samples were imaged with a 561-nm excitation laser and a 570–620 nm acquisition window. DAPI-labelled samples were imaged with a 405 nm excitation laser and a 430–490 nm acquisition window.

Immunostained muscle image analysis. Spatial, ratiometric analysis of the distribution of complex IV and V within the muscle was performed. First, binary image masks were created by thresholding the 3D image stacks for both the complex IV and V signals, thereby allowing the removal of all pixels with negligible signal. Thus, ratiometric analysis proceeded only for pixels where signal could be detected. Thresholding was performed in ImageJ using Li's Minimum Cross Entropy thresholding method. Ratio images were created by dividing the complex IV signal by the complex V signal for all remaining pixels. The look-up table for the ratiometric images was set as follows: the mean of the middle 80% of the data and removed pixels were set to black; green intensity was linearly increased from 0 at the mean up to 255 at 2× mean and above; red intensity was linearly increased from 0 at the mean up to 255 at 0.5× mean and below. The look-up table is included in Fig. 4a for reference. The spatial distribution of green (relatively higher complex IV) and red (relatively higher complex V) pixels was assessed for the PVM and IFM pools by tracing the outer boundary of the fibre and tracing the inner boundary of the PVM pool, allowing separate analysis of each region. The total intensity of green pixels was divided by the total intensity of red pixels for each region. This analysis was performed on 12 total fibres from 5 mice: 6 fibres in cross-section and 6 in longitudinal section.

Additionally, the fluorescent intensity profiles of the complex IV and V signals were assessed individually with respect to their distance from the fibre boundary. The background fluorescence for each channel was determined as the mean of a manually drawn ROI outside the fibre of interest where no other fibre was present. Background was subtracted from all pixels in each image. A binary mask of the fibre was created by manually tracing the fibre of interest, and a distance map from the fibre boundary was created using ImageJ. This distance map allowed for the selection of all pixels in the complex IV and V images that were a given distance from the fibre boundary and the mean value for each distance calculated. This was performed for both the complex IV and complex V signals individually. For a given image, all values for the complex IV intensity profile were divided by the maximal complex IV signal measured near the fibre boundary and all values for the complex V intensity profile were divided by the maximal complex V signal measured near the fibre boundary. This analysis was performed on images of 4 fibres from 4 mice.

Clear native PAGE and western blotting. Mouse skeletal muscle mitochondria were used to test the specificity of the primary antibodies used in the whole muscle immunostaining studies. Skeletal muscles from the hindlimbs of two mice (~2.5 g) were dissected and combined for mitochondrial isolation as described previously[32] except omitting the Percoll gradient and subsequent washes. Cytochrome *a*,a_3 (cyt *a*) content was determined as described previously[33]. Clear Native (CN) PAGE was performed using the NativePAGE Novex Bis-Tris System (Invitrogen, Carlsbad, CA) with 4–16% 1 mm bis-tris gels and the anode and cathode buffers for high resolution CN PAGE 2 (hrCNE2)[34]. Mitochondria (10 pmol cyt *a*) were solubilized with 3% w/v dodecyl maltoside and added to each lane and run at 4 °C for 1 h at 150 V and 1.3 h at 250 V. Proteins were transferred to a PVDF membrane using a XCELL II blot module (Invitrogen, Carlsbad, CA) at 25 V for 3 h at 4 °C. After transfer, the PVDF membrane was incubated in 8% acetic acid for 15 min, rinsed with water, placed in Pierce StartingBlock buffer (ThermoScientific, Rockford, IL) for 1 h at room temperature, and incubated in the primary antibodies used above (1:1,000 in StartingBlock) overnight at 4 °C. The membrane was then washed for 20 min in StartingBlock buffer and two more times for 20 min in PBS + 0.05% Tween 20. Incubation with secondary antibodies (Alexa 488 donkey anti-mouse IgG and Alexa 488 goat, anti-rabbit IgG, Abcam 150105 and 150077, respectively) was done at a 1:10,000 dilution in PBS + Tween 20 for 1.5 h at room temperature in the dark. Membranes were washed 2× in PBS + Tween 20 for 20 min in the dark and imaged on a Typhoon variable mode imager.

Single muscle fibre isolation and imaging. For these studies, mouse soleus muscle fibres were used to increase the probability of acquiring mitochondria-rich muscle fibres. The basic incubation medium (IM) was composed of (in mM): NaHEPES (10), NaCl (137), KCl (5.4), $CaCl_2$ (1.8), $MgCl_2$ (0.5), $NaPO_4$ (0.5), glucose (10), NaPyruvate (1) and butanedione monoxomine (BDM, 20). The soleus muscles were quickly isolated tendon to tendon. The intact muscle was then placed in the IM solution containing 3 mg ml^{-1} collagenase D (Roche) for 30 to 45 min in a shaking water bath at 37 °C. The tissue was removed from the collagenase solution and placed in the IM solution alone. The fibres were gently teased from the muscle bundle and long, non-contracted fibres were used for study. For TMRM labelling, the fibres were incubated in IM containing 5 nM TMRM for >20 min. For regional uncoupling of mitochondria, MitoPhotoDNP[25], a gift from S. Caldwell and R. Hartley (University of Glasgow), was used. In this preparation, we found that we needed an incubation concentration of 20 μM for approximately 30 min at room temperature to achieve adequate loading of the mitochondria. The loaded cells with TMRM and MitoPhotoDNP still in the IM were transferred to a Petri dish with a no. 1.5 coverglass bottom (MatTek) treated with Cell-Tak (BD Bioscience) for microscopy.

Confocal microscopy was conducted on an inverted ZEISS 780 with a 40×, 1.2 NA water objective. 976 × 976 pixel, 12-bit images of 106.3 × 106.3 μm were collected while TMRM was excited with a 561 nm laser at 0.2% power. Emitted light was collected from 570 nm to 695 nm. To photoactivate MitoPhotoDNP, an ROI was drawn along the longitudinal axis within the centre of the cell and exposed to UV light from a 355 nm laser (Coherent Inc.) at 40% power for 10 iterations. It is important to note that the mild depolarization used here is of similar magnitude to the full change in membrane potential of mitochondria moving from resting to maximally active states (so called State 4 to State 3 transition) where the ~200 mV membrane potential decreases by only 10–20%. To be physiologically relevant, any conductive mechanisms must be responsive to small changes in membrane potential. In control experiments, fibres which were not loaded with MitoPhotoDNP were exposed to an identical dose of UV light in an ROI in the centre of the cell. Experiments were only performed on fibres with a strong and homogeneous TMRM signal as low and heterogeneous signals are potential signs of a damage caused by the isolation process.

Quantitative image analysis was performed using ImageJ on an average of five images taken before (Pre) and of five images taken immediately after UV exposure/MitoPhotoDNP release (Post). Image acquisition time varied slightly depending on cell width but averaged ~200 ms per frame. Both Pre and Post images were amplitude thresholded to separate the mitochondrial and cytosolic+nuclear TMRM signals. Cells were then analysed by cell region: paravascular, intra-fibrillar and irradiated. The irradiated region is the area in the centre of the cell exposed to UV light, and the boundary between the paravascular and intra-fibrillar regions is clearly distinguishable due to the regularity of the low cytosolic TMRM signal in the intra-fibrillar region where myofibrils are located. The mean TMRM signal was then measured for all mitochondrial pixels and all cytosolic+nuclear pixels in each cell region in both the Pre and Post images. From these values, a Post/Pre ratio was calculated for the mitochondria and cytosol+nuclei in each cell region. The ratio between the mitochondrial and cytosolic+nuclear TMRM intensities (mito/cyto ratio) was also calculated for each cell region in the Pre and Post images as this is a more valid measure of mitochondrial membrane potential than just the mitochondrial signal alone.

Live cell mitochondrial morphology. Super-resolution microscopy was also performed on the live, isolated fibres loaded with TMRM using stimulated emission depletion (STED) methodology. Although the spatial resolution of confocal microscopy is theoretically sufficient to measure mitochondrial cross-sectional diameters (mean values >300 nm) in these cells, the highly dense nature of the PVM pool makes it difficult to resolve the spaces between mitochondria. Thus, the improved spatial resolution with STED provides better definition of individual PVM structures. Gated STED images were obtained using a commercial Leica SP8 STED 3X system (Leica Microsystems, Mannheim, Germany), equipped with a white light excitation laser. A 100×/1.4 NA oil immersion objective lens (HCX PL APO STED white, Leica Microsystems) was used for imaging. Using TMRM as the mitochondrial probe for STED allowed the measurement of mitochondrial structure under the same conditions used to assess mitochondrial membrane potential conduction. Further, TMRM is highly recommended by Leica for live cell STED as it is very bright and highly photostable, which is supported by the results of the UV irradiation experiments described above. Also, tetramethylrhodamine has been used previously to collect STED images in live cells[35]. Image stacks were collected as 8 bit, 1,024 × 1,024 images with 30 nm *x* and *y* pixel sizes, 6 line averages and 0.2 μm *z*-steps. Muscle cells were imaged with 550 nm excitation light, a 660 nm STED depletion laser at 20% power and an emission window from 564 nm to 640 nm. Tandem confocal images (no STED depletion, 3× lower excitation power) were also obtained with confocal and STED images collected sequentially, frame by frame, with the confocal image first.

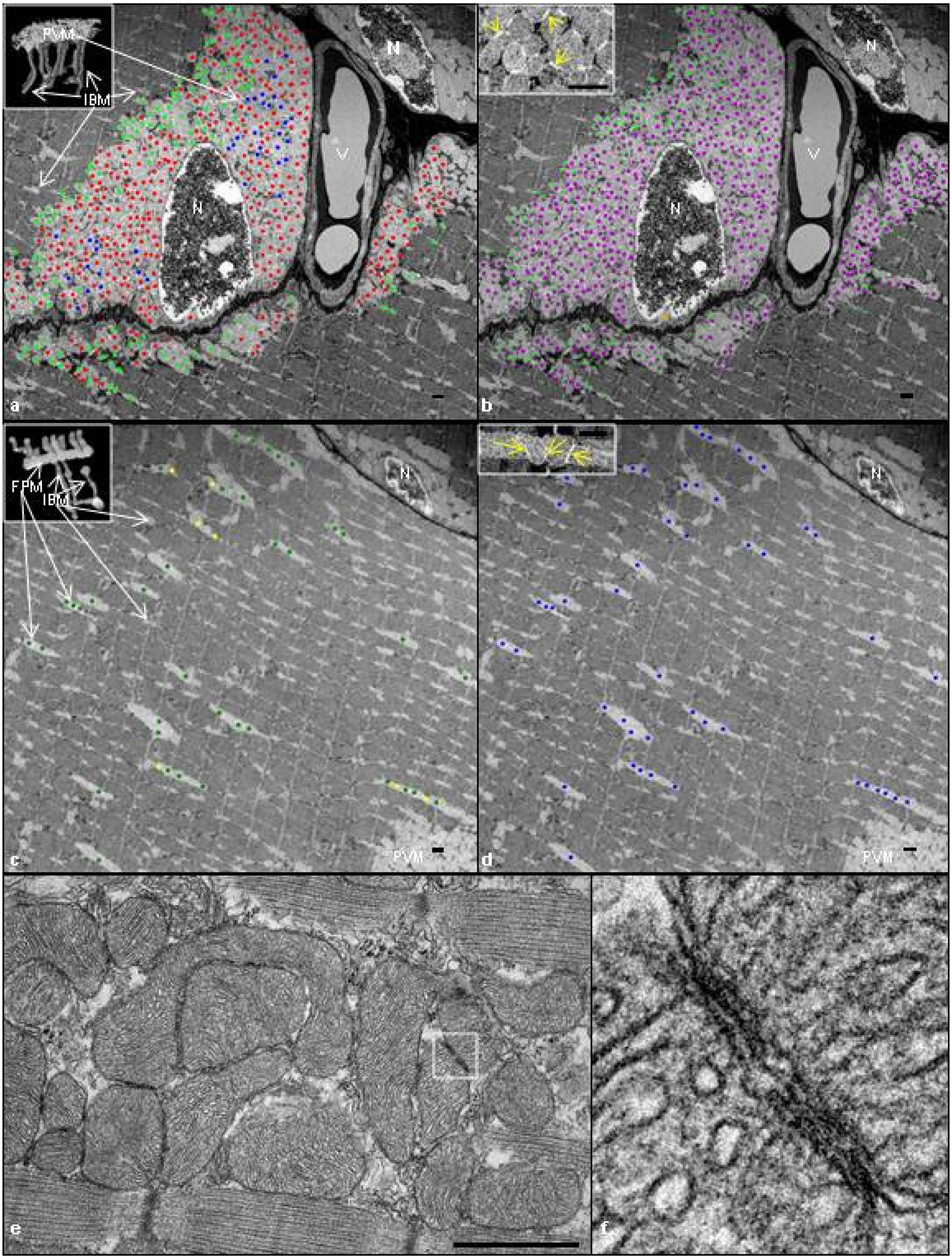
Measurement of mitochondrial diameters was performed in ImageJ using the raw STED image stacks. A line was drawn across the maximal cross-sectional diameter and the length determined as the full width at half maximum (FWHM) of the linear fluorescent intensity profile. We analysed 175 mitochondria from 13 muscle fibres from 3 animals (see Extended Data Table 1). STED image stacks were deconvolved using Huygens software and the CMLE algorithm for presentation purposes only.

Fast Fourier transform (FFT) IBM analysis. In the isolated muscle fibres, IBM were generally spaced every ~2 μm adjacent to the *z*-lines in the isolated cells. Since depolarization dissipates the TMRM signal from the mitochondria to the cytosol, we reasoned that an FFT of the TMRM signal along the long axis of the cell would selectively highlight, due to its periodicity, the IBM pool signal alterations. To accomplish this, a large ROI of the intra-fibrillar region was selected, eliminating the PVM pool. The FFT were conducted serially along the long axis of the cell summing the power spectrum of each FFT to provide the overall power spectrum of the ROI. The IBM represented a unique resonance in the 0.5 μm^{-1} region permitting the specific analysis of the IBM amplitude using this approach.

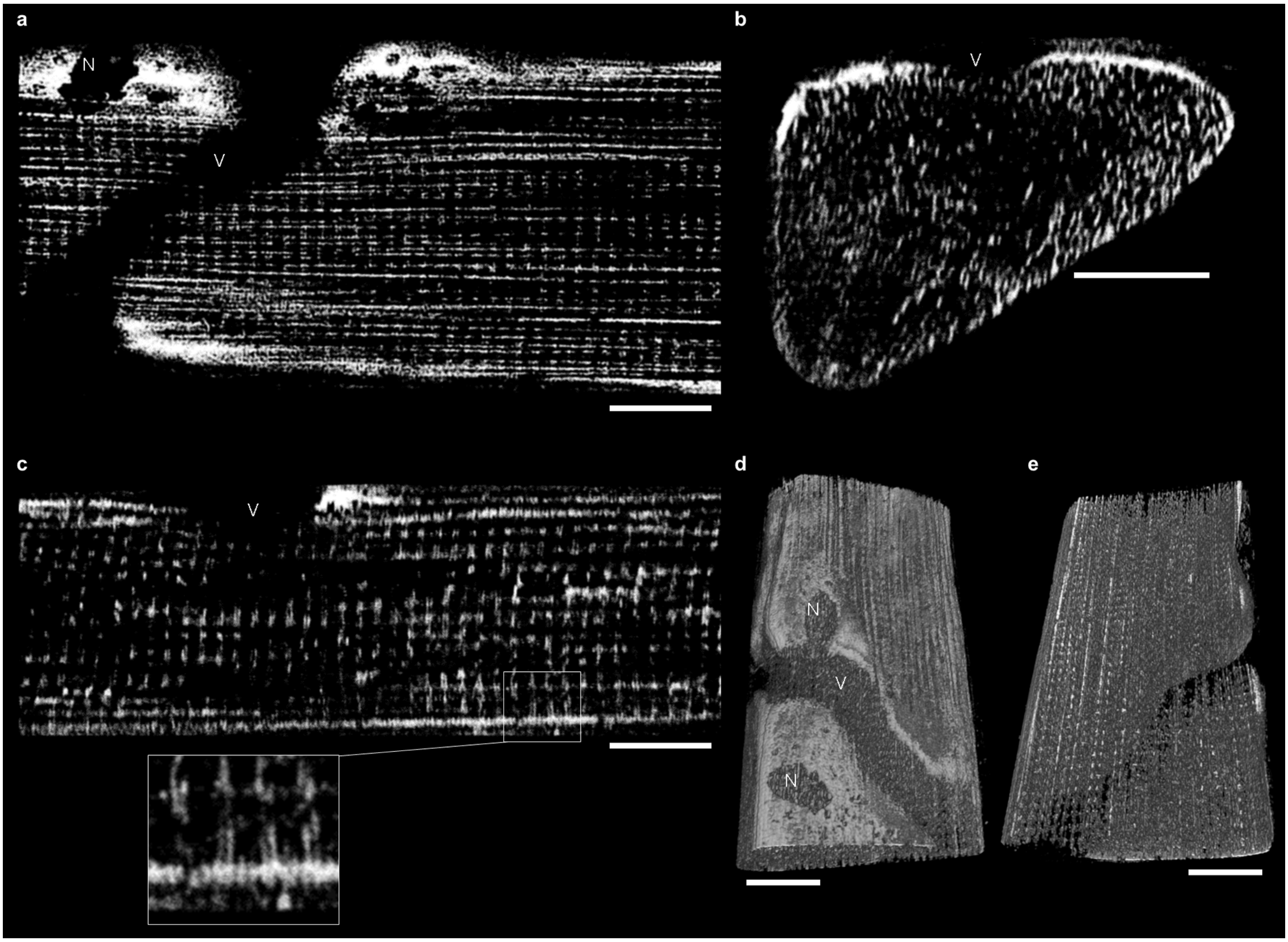
Statistical analyses. Differences in mitochondrial diameter for the four mitochondrial pools were assessed using a one-way ANOVA with a Tukey's HSD post hoc test and a *P* value of 0.05. Differences in the relative distribution of complex IV and complex V between the PVM and IFM pools were determined using paired *t*-tests and a *P* value of 0.05. Differences between isolated muscle fibres before and after UV exposure and for different cell regions were assessed by ANOVA with a Tukey's HSD post hoc test and a *P* value of 0.05.

28. Bakalar, M. *et al.* Three-dimensional motion tracking for high-resolution optical microscopy, *in vivo*. *J. Microsc.* **246,** 237–247 (2012).
29. Narayan, K. *et al.* Multi-resolution correlative focused ion beam scanning electron microscopy: applications to cell biology. *J. Struct. Biol.* **185,** 278–284 (2014).
30. Murphy, G. E. *et al.* Correlative 3D imaging of whole mammalian cells with light and electron microscopy. *J. Struct. Biol.* **176,** 268–278 (2011).
31. Kremer, J. R., Mastronarde, D. N. & McIntosh, J. R. Computer visualization of three-dimensional image data using IMOD. *J. Struct. Biol.* **116,** 71–76 (1996).

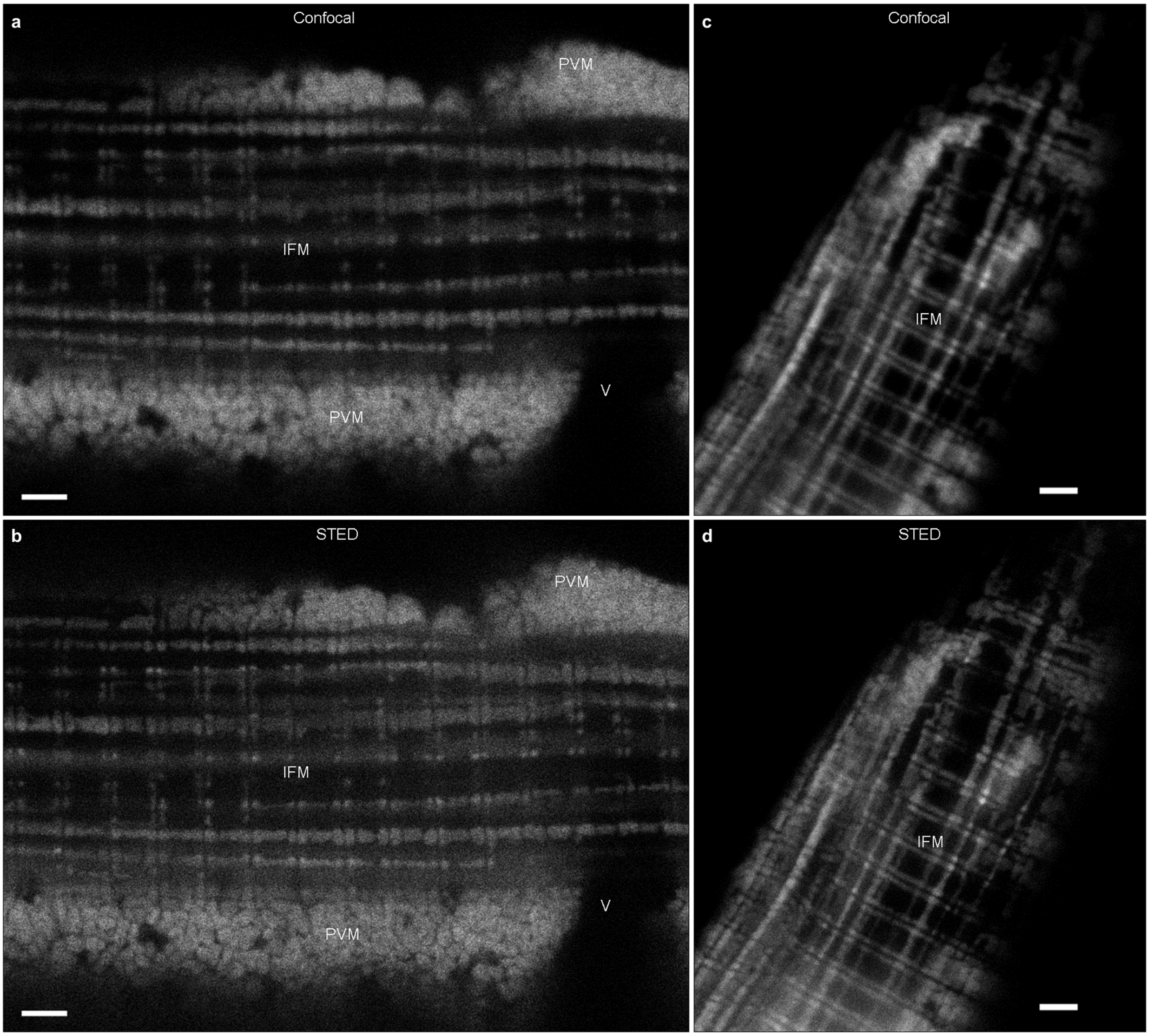
32. Glancy, B. & Balaban, R. S. Protein composition and function of red and white skeletal muscle mitochondria. *Am. J. Physiol. Cell Physiol.* **300,** C1280–C1290 (2011).
33. Balaban, R. S., Mootha, V. K. & Arai, A. Spectroscopic determination of cytochrome c oxidase content in tissues containing myoglobin or hemoglobin. *Anal. Biochem.* **237,** 274–278 (1996).
34. Wittig, I., Karas, M. & Schagger, H. High resolution clear native electrophoresis for in-gel functional assays and fluorescence studies of membrane protein complexes. *Mol. Cell. Proteomics* **6,** 1215–1225 (2007).
35. Hein, B. *et al.* Stimulated emission depletion nanoscopy of living cells using SNAP-Tag fusion proteins. *Biophys. J.* **98,** 158–163 (2010).



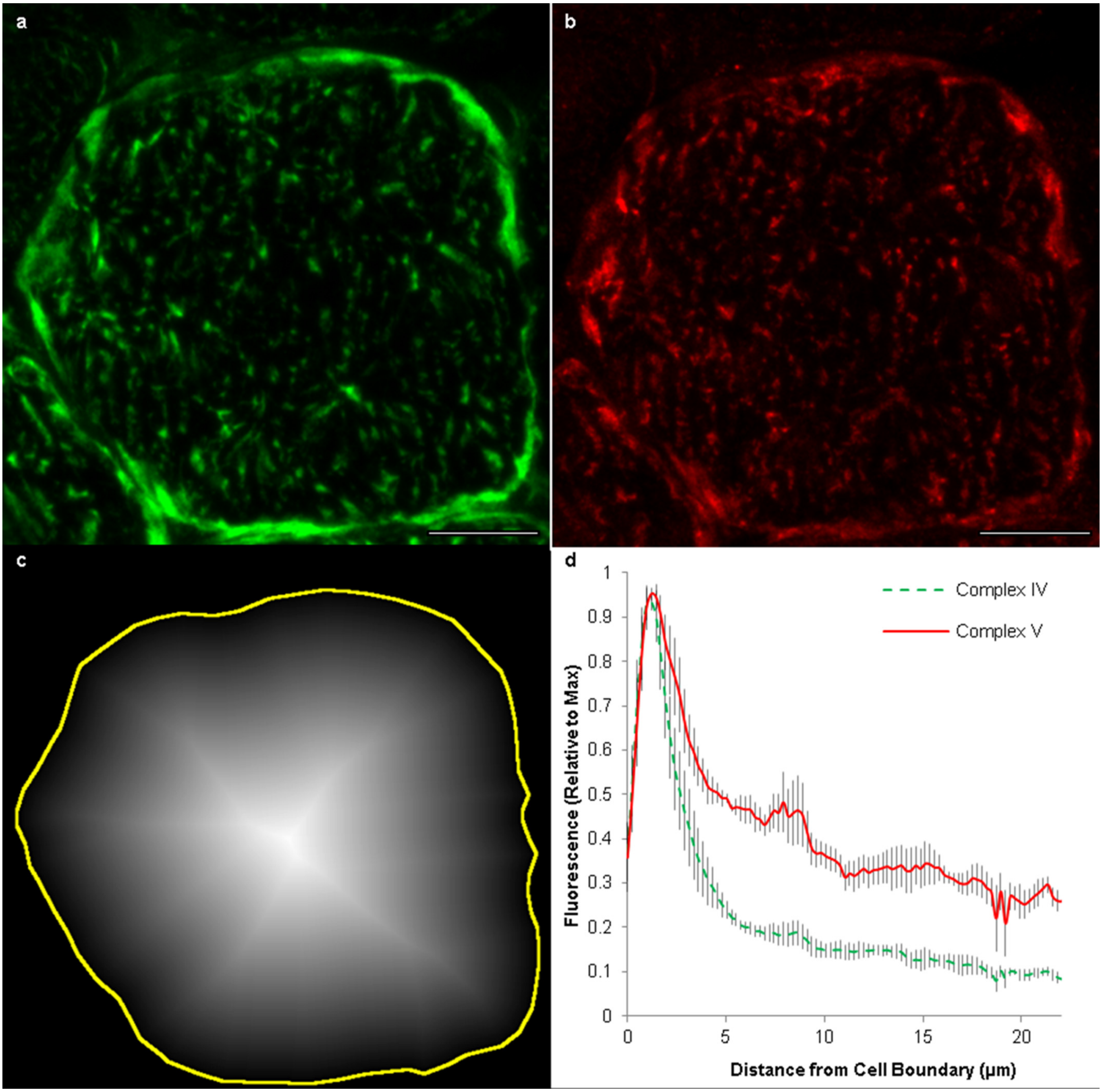
Extended Data Figure 1 | Mitochondrial coupling assessment. **a**, **b**, Single image from the middle of a FIB-SEM muscle volume including two nuclei (N) and a blood vessel (V). **a**, Direct coupling. PVM tagged green are those where a single mitochondrion projects from a PVM into an IBM with a continuous outer membrane (see inset) within the imaged volume. Red-tagged PVM did not project into IBM. Blue-tagged PVM continued out of the field of view and could not be classified. **b**, Electron-dense contact site (EDCS) coupling. Magenta-tagged mitochondria were connected by EDCS (see yellow arrows in inset). Additional green dots were added to PVM which project into IBM (see **a**). Greater than 99% of tracked PVM were coupled by EDCS. **c**, **d**, Single image from the middle of the intra-fibrillar region of a FIB-SEM volume. **c**, Direct coupling. FPM tagged green are those where a single mitochondrion projects from an FPM into an IBM with a continuous outer membrane (see inset). Yellow-tagged FPM did not project into IBM. **d**, EDCS coupling. FPM tagged blue were connected to an adjacent FPM through an EDCS (see yellow arrows in inset). Images are representative of 4 samples with significant paravascular and intra-fibrillar mitochondrial content from 4 animals. **e**, A longitudinal TEM mouse tibialis anterior muscle image showing the close association between adjacent mitochondria. **f**, Close up view of an EDCS highlighted in **e** showing the convergence of mitochondrial membranes between two adjacent mitochondria. Scale bars, 750 nm.



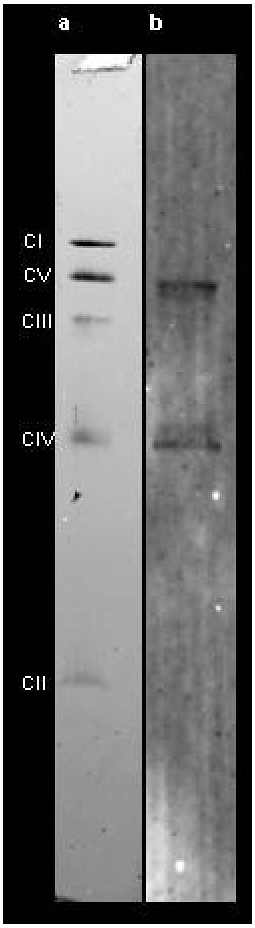
Extended Data Figure 2 | Multi-photon microscopy (MPM) images of fresh muscle fibres *in situ*. **a**, Endogenous mitochondrial NAD(P)H signal from a single *XY* image within a 3D volume of a muscle fibre. In this orientation, paravascular mitochondria (PVM) are apparent as clusters around the embedded capillary (V) and nuclei (N). Fibre parallel mitochondria (FPM) are seen as horizontal lines while I-band mitochondria (IBM) are seen as discrete spots. There are very few vertical lines in this image due to the infrequency of cross-fibre connection mitochondria (CFCM). The full volume can be seen in Supplementary Video 3. **b**, A *YZ* image from the same fibre volume as in **a**. PVM appear lateral to the embedded capillary (V), while IBM appear as vertical lines and FPM are seen as discrete spots. Note the lack of horizontal lines (CFCM) in this image or the accompanying video (Supplementary Video 4). **c**, An *XZ* image of the same muscle volume as in **a** and **b**. PVM are located on the cell periphery; IBM are seen as vertical lines; FPM are seen as horizontal lines; and CFCM as discrete spots. In the inset, pairs of IBM projecting out from PVM are highlighted. **d**, A 3D rendering of the endogenous mitochondrial NAD(P)H signal within the muscle fibre shown in **a**–**c**. 360° views of this 3D rendering are shown in Supplementary Video 5. **e**, A view of the interior of the 3D rendering from **d** showing the regularity of the mitochondrial network within skeletal muscle. The field of view for the muscle volume in this image is 102.4 × 51.2 × 36.8 μm in *x*, *y* and *z*, respectively. Scale bars, 15 μm. Mean fibre volume from these MPM images was 159,431 ± 15,507 μm^3. Images are representative of 4 fibres from 3 mice.



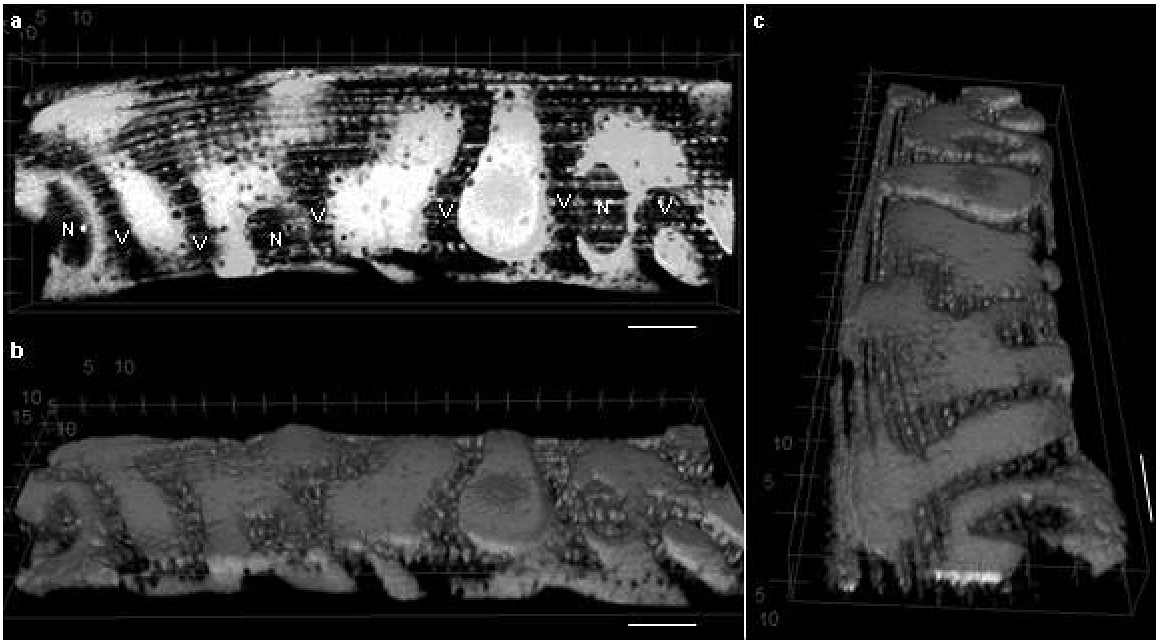
Extended Data Figure 3 | Super resolution microscopy allows improved visualization of individual mitochondria. **a**, Single confocal microscopy image of an isolated muscle fibre loaded with mitochondrial membrane potential dye, TMRM. PVM, paravascular mitochondria; IFM, intra-fibrillar mitochondria; V, vessel. **b**, Single stimulated emission depletion (STED) microscopy image of the same fibre as in **a**. **c**, The average of a confocal microscopy image stack of a TMRM loaded isolated muscle fibre. **d**, The average of a STED microscopy image stack of the same fibre as in **c**. For **c** and **d**, image volume was 24.30 × 27.47 × 1.75 μm in *x*, *y* and *z*, and images shown are the average of 9 images taken with 219 nm *z*-steps. Confocal images were acquired before STED images at each image depth. All images were acquired with *x* and *y* pixel sizes of 30 nm. Scale bars, 2 μm. As most muscle mitochondria are larger than 300 nm, the improved resolution with STED is primarily noted by the increased clarity of the spaces between individual mitochondria, particularly the PVM. Images are representative of 13 fibres from 3 mice.


a
b
c
d
Complex IV
Complex V
Fluorescence (Relative to Max)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
5
10
15
20
Distance from Cell Boundary (µm)

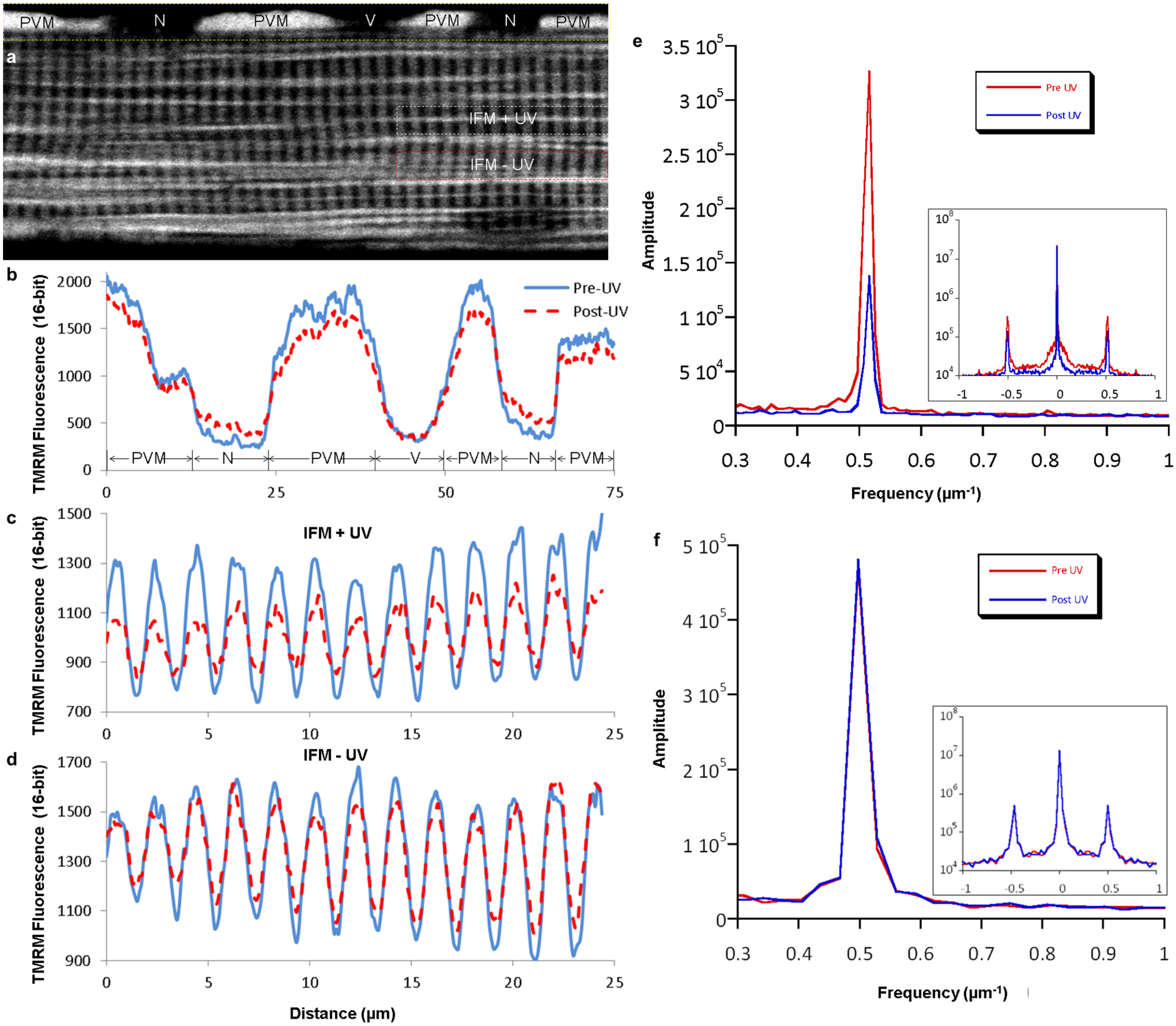
Extended Data Figure 4 | Complex V expression is relatively higher than complex IV in the muscle fibre interior. **a**, Representative confocal microscopy image of a cross-section of a muscle fibre immunostained for complex IV. **b**, The same muscle section depicted in **a** immunostained for complex V. Scale bars, 15 µm. All images analysis was performed on the raw images. Brightness and contrast have been increased for the images shown here to improve presentation. **c**, Distance map from the boundary of the cell (yellow line) towards the interior of the fibre where image intensity corresponds to distance from the cell boundary. The mean fluorescence signal for all pixels of a given distance from the cell boundary was assessed for both complex IV and complex V. Images representative of 4 fibres from 4 mice. **d**, Results from intensity profile analysis of 4 muscle fibres from 4 animals show relatively increased complex V in the fibre interior or, conversely, relatively higher complex IV near the fibre periphery. The red, solid line shows the mean intensity profile for complex V from the outside to the inside of the fibre. The green, dotted line shows the mean intensity profile for complex IV from the outside to the inside of the fibre. Values are normalized to the maximal intensity from each intensity profile to allow comparison of images of varying intensities. Grey shading around each line shows the standard error.



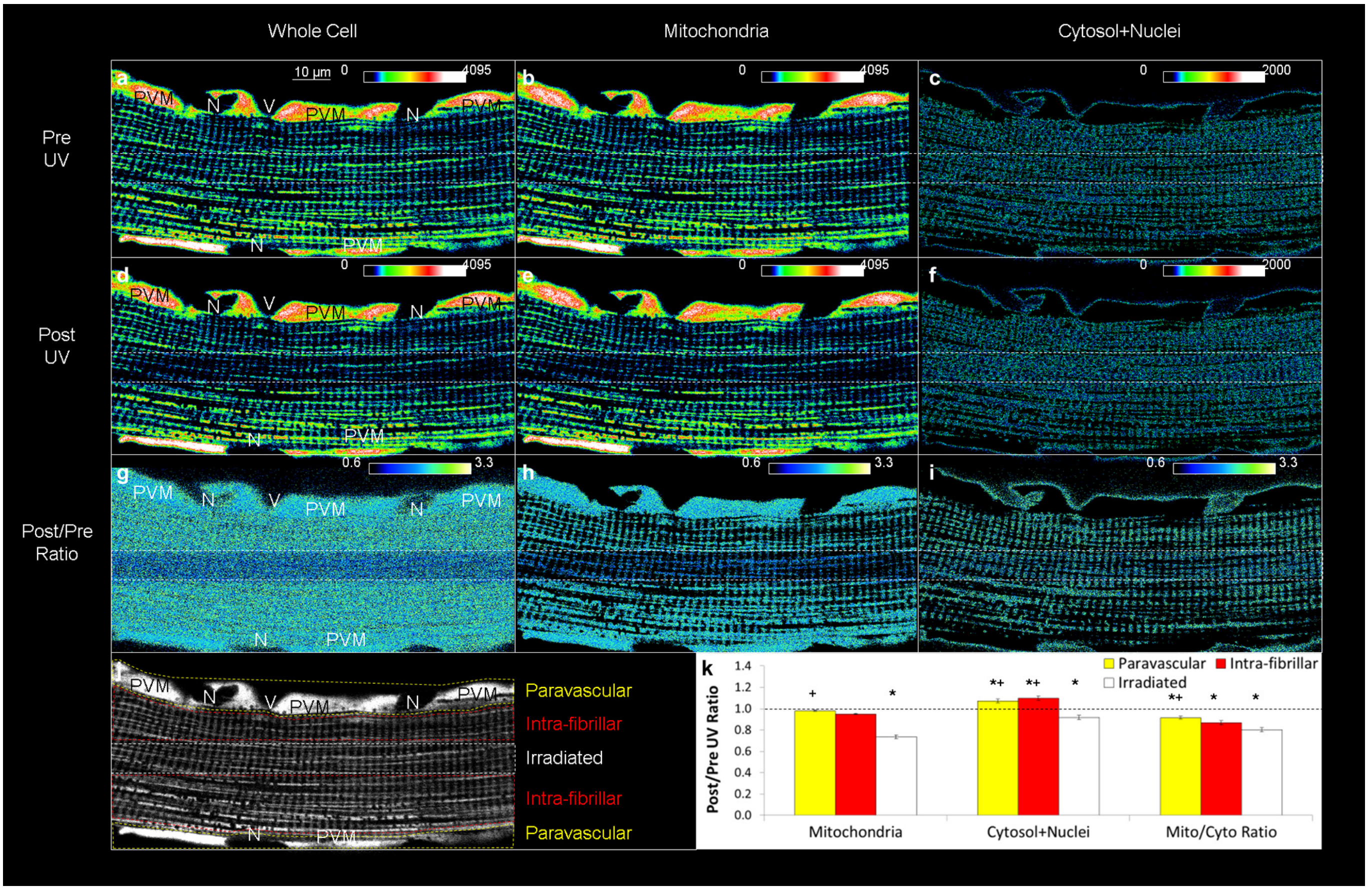
Extended Data Figure 5 | Primary antibody specificity. **a**, Coomassie-stained CN-PAGE gel after transferring isolated mouse skeletal muscle mitochondrial proteins to PVDF membrane for western blotting. Mitochondrial oxidative phosphorylation complexes I–V can be resolved as individual bands. **b**, Western blot dual immunostained for complex V, β-subunit (upper band), and complex IV, subunit I (lower band), with the same primary antibodies as used for the muscle section dual immunostaining as shown in Fig. 3 and Extended Data Fig. 4.



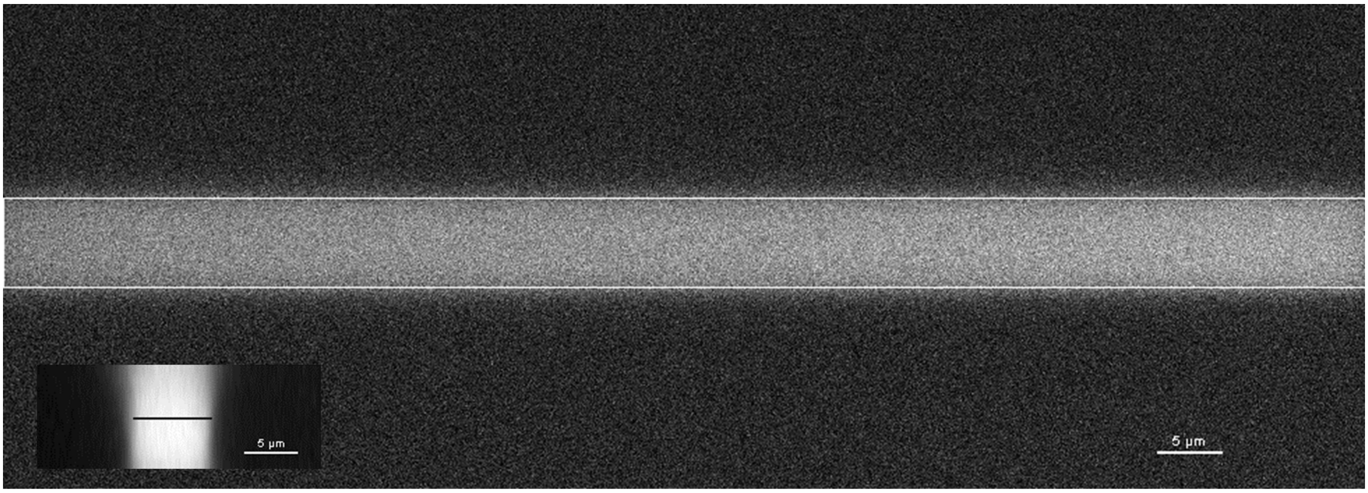
Extended Data Figure 6 | Sarcolemmal capillary grooves remain even after removal of capillaries. **a**, Longitudinal view of a 3D rendering of the outer 10 μm of an enzymatically isolated muscle fibre imaged as the mitochondrial TMRM signal by confocal microscopy. Blood vessels (V) and nuclei (N) are apparent as negative signals. **b**, **c**, Rotation of the volume in **a** around the longitudinal (**b**) and cross-sectional (**c**) axes reveals the grooves in the sarcolemma where capillaries were once present. Scale bars, 10 μm. Images representative of 11 fibres from 4 mice.



Extended Data Figure 7 | Additional analyses of mitochondrial membrane potential conduction. **a**, Confocal image of an isolated muscle fibre loaded with TMRM showing regions chosen for representative line profile analysis. Images are representative of 11 fibres from 4 mice. **b**, Line profile of TMRM fluorescence in the PVM+nuclear (N) region of the cell (marked by yellow dotted lines in **a**) before and after activation of MitoPhotoDNP shows a decrease in PVM signal, increase in N signal, and no change outside of the cell where a blood vessel (V) was located before enzymatic fibre isolation. **c**, Line profile of TMRM fluorescence in the region of the cell irradiated with UV light (irradiated, marked by white dotted lines in **a**) showing a decrease in mitochondrial signal (peaks) and an increase in cytosolic signal (troughs) after MitoPhotoDNP activation. **d**, Line profile of TMRM fluorescence in the non-irradiated IFM region of the cell (intra-fibrillar, marked by red dotted lines in **a**) showing a decrease in mitochondrial signal (peaks) and an increase in cytosolic signal (troughs) after MitoPhotoDNP activation. **c** and **d** take advantage of the regular pattern of IBM and cytosol in the intra-fibrillar space. However, any FPM in the regions selected for analysis may confound these results. A more robust analysis that selects only the IBM is shown in **e** and **f**. **e**, Representative fast Fourier transform (FFT) power spectrum of IFM TMRM signal reveals a 61 ± 3% decrease ($n = 9$ experiments) in the amplitude of IBM component (0.5 μm^{-1} frequency) after MitoPhotoDNP activation. Inset: log scale full FFT power spectrum. **f**, Representative FFT analyses of IFM mitochondria reveals no change (pre/post = 1.0 ± 0.03, $n = 5$ experiments) in the TMRM amplitude of IBM (0.5 μm^{-1} frequency) after UV exposure when MitoPhotoDNP is not present. Inset: log scale full FFT power spectrum.

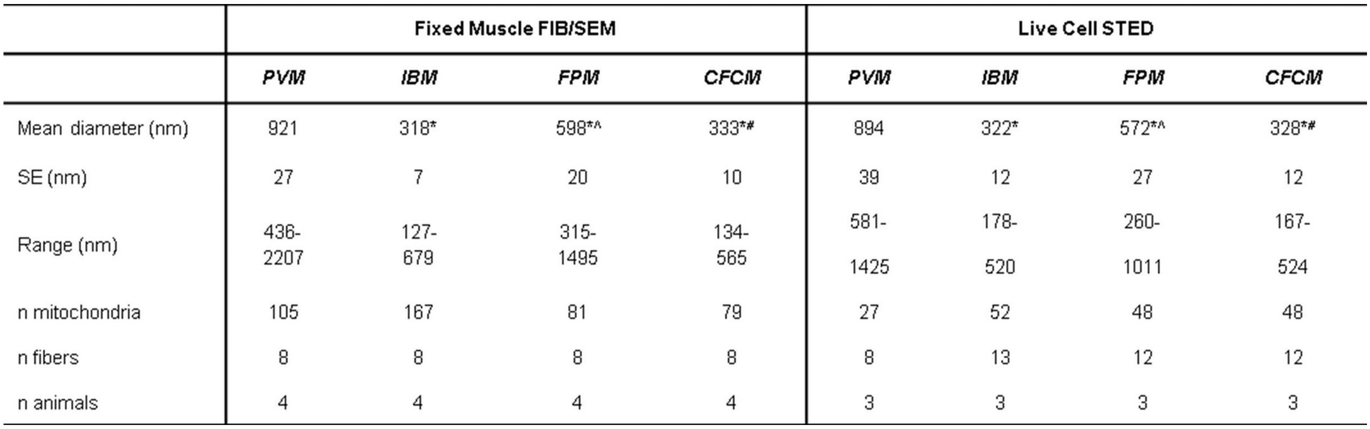


Extended Data Figure 8 | Effect of UV exposure on membrane potential without MitoPhotoDNP. **a**, Representative heat map image of an isolated mouse soleus muscle fibre loaded with 5 nM TMRM. N, nucleus; V, vessel. **b**, **c**, Image from **a** thresholded to include only the TMRM signal in the mitochondria (**b**) or cytosol+nuclei (**c**). **d**–**f**, Whole cell (**d**), mitochondrial (**e**) and cytosolic+nuclear (**f**) TMRM signals immediately after UV exposure in the centre region of the cell (outlined by the white dotted lines). **g**–**i**, Post/pre ratiometric whole-cell (**g**), mitochondrial (**h**), and cytosolic+nuclear (**i**) TMRM images showing little change outside of the UV-exposed area. **j**, Greyscale image of **a** highlighting the different cell regions used for quantitative analysis. All pixels were analysed. **k**, TMRM signal decreased in both the mitochondria and cytosol in the UV-exposed region indicative of slight photobleaching. Images are representative of 10 fibres from 3 mice. Data represent the mean ± s.e. from 10 experiments from 3 animals. Asterisk indicates significantly different (ANOVA, $P < 0.05$) from a post/pre ratio of 1.0. Plus symbol indicates significantly different ($P < 0.05$) from IFM+UV region. See Supplementary Video 10 for time course video.



Extended Data Figure 9 | UV light was restricted to the drawn ROI. Difference *XY* image of a uniformly blue fluorescent slide before and after exposure to UV light in the drawn ROI (white outline). Inset: *YZ* view shows that the UV light was largely maintained to the ROI drawn (black line). Scale bars, 5 µm. Images are representative of 3 experiments.

Extended Data Table 1 | Skeletal muscle mitochondrial cross-sectional diameters



	Fixed Muscle FIB/SEM				Live Cell STED			
	PVM	*IBM*	*FPM*	*CFCM*	*PVM*	*IBM*	*FPM*	*CFCM*
Mean diameter (nm)	921	318*	598*^	333*#	894	322*	572*^	328*#
SE (nm)	27	7	20	10	39	12	27	12
Range (nm)	436- 2207	127- 679	315- 1495	134- 565	581- 1425	178- 520	260- 1011	167- 524
n mitochondria	105	167	81	79	27	52	48	48
n fibers	8	8	8	8	8	13	12	12
n animals	4	4	4	4	3	3	3	3

FIB-SEM, focused ion beam scanning electron microscopy; STED, stimulated emission depletion microscopy; PVM, paravascular mitochondria; IBM, I-band mitochondria; FPM, fibre parallel mitochondria; CFCM, cross fibre connection mitochondria. Statistical test was analysis of variance ($P < 0.05$). There were no significant differences between FIB-SEM and STED. Diameters were measured at the maximum cross-section of the short axis of the mitochondria which were roughly ellipsoidal or cylindrical in shape. STED diameters were determined by the full-width half-maximum (FWHM) of the linear fluorescence intensity profile, and FIB-SEM diameters were determined as the distance of a line drawn across a mitochondrion.

* Denotes significantly different from PVM.

^ Denotes significantly different from IBM.

# Denotes significantly different from FPM.

doi:10.1038/nature14482

Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex

Lanfeng Wang[1]*, Yu Zhou[2]*, Liang Xu[1]*, Rui Xiao[2]*, Xingyu Lu[3], Liang Chen[2], Jenny Chong[1], Hairi Li[2], Chuan He[3], Xiang-Dong Fu[2] & Dong Wang[1]

DNA methylation at selective cytosine residues (5-methylcytosine (5mC)) and their removal by TET-mediated DNA demethylation are critical for setting up pluripotent states in early embryonic development[1,2]. TET enzymes successively convert 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC), with 5fC and 5caC subject to removal by thymine DNA glycosylase (TDG) in conjunction with base excision repair[1–6]. Early reports indicate that 5fC and 5caC could be stably detected on enhancers, promoters and gene bodies, with distinct effects on gene expression, but the mechanisms have remained elusive[7,8]. Here we determined the X-ray crystal structure of yeast elongating RNA polymerase II (Pol II) in complex with a DNA template containing oxidized 5mCs, revealing specific hydrogen bonds between the 5-carboxyl group of 5caC and the conserved epi-DNA recognition loop in the polymerase. This causes a positional shift for incoming nucleoside 5′-triphosphate (NTP), thus compromising nucleotide addition. To test the implication of this structural insight *in vivo*, we determined the global effect of increased 5fC/5caC levels on transcription, finding that such DNA modifications indeed retarded Pol II elongation on gene bodies. These results demonstrate the functional impact of oxidized 5mCs on gene expression and suggest a novel role for Pol II as a specific and direct epigenetic sensor during transcription elongation.

Epigenetic DNA methylation (5mC) is an important regulator of gene transcription recognized by several families of protein readers, such as methyl-CpG-binding domain proteins (MBDs) and ubiquitin-like PHD and RING finger domain-containing proteins (for example, UHRF1), and certain zinc-finger proteins (kaiso; also known as ZBTB33)[9,10]. TET enzymes iteratively oxidize 5mC to 5hmC, 5fC, and 5caC[3–6], and TDG coupled with base excision repair further process 5fC/5caC to complete DNA demethylation (Fig. 1a)[5,6]. An open question is whether 5fC and 5caC are simple DNA demethylation intermediates or have active roles in gene expression.

Genomic mapping revealed specific enrichment of 5fC and 5caC at enhancers, promoters and gene bodies[7,8]. Moreover, a number of protein complexes involved in transcription, splicing, chromatin remodelling and DNA repair have been identified to selectively bind synthetic DNA oligonucleotides containing oxidized 5mCs (oxi-mCs)[11–13]. A previous study by our group indicates that these modifications induce transient pausing of purified yeast and mammalian RNA polymerase II elongation complex (Pol II EC) *in vitro*[14]. Together, these observations imply that different oxi-mCs may influence gene expression[15]. Importantly, our work suggests that Pol II has the capacity to directly sense the demethylation state (oxi-mCs) of template DNA during transcription.

To understand the molecular basis underlying the Pol II EC recognition of oxi-mCs, we performed structural studies of the complex assembled on an RNA/DNA scaffold that contains a 5caC at the i+1 site (Fig. 1b) to mimic the stage when Pol II EC encounters 5caC during transcription elongation. This scaffold recapitulated the impediment of Pol II elongation in the *in vitro* reactions (Fig. 1c)[14]. The crystal structure (EC-I) revealed that the upstream RNA/DNA hybrid region maintains a post-translocation state register in which the active site is empty and ready for NTP loading (Fig. 1d). About 50% of 5caC nucleobase (yellow coloured in Fig. 1d, h, see also Extended Data Fig. 1a, b) accommodates at a new translocation intermediate position, located about halfway between the canonical i+1 and i+2 sites. The other 50% of 5caC nucleobase is partially inserted into the i+1 position (cyan coloured in Fig. 1d, g). Importantly, we detected specific hydrogen bonds between the 5-carboxyl moiety of 5caC and the side chain of residue Q531 at a loop in the fork region of Rpb2 (the second largest subunit) (Fig. 1e, f)[16]. We termed it the 'epi-DNA recognition loop' or 'fork loop 3', because it recognizes the epigenetic DNA modification in the major groove and is next to the previously identified fork loop 1 and fork loop 2 within the fork region[16]. The specific hydrogen-bonding interactions with 5caC result in a 90-degree rotation of the side chain of Q531 of the yeast Pol II Rpb2, switching its interacting partner in the upstream RNA/DNA hybrid region[17] to the nucleobase of 5caC at i+1 position register (Fig. 1e, f). This causes 5caC to shift into a new translocation intermediate position right above the bridge helix (Fig. 1e, f, h), which we termed the 'midway position'.

To investigate the potential impact of 5caC on nucleotide incorporation, we next solved the structure of the Pol II EC with a 5caC at the i+1 site in the presence of a non-hydrolysable GTP analogue (GMPCPP) to mimic the state of GTP binding opposite 5caC (EC-II). We found that while 5caC forms a canonical Watson–Crick base pair with GMPCPP (Fig. 2a and Extended Data Fig. 1c, d), the base pair shifts to another translocation intermediate position, ~1.5 Å away from its canonical position towards the downstream main channel (Fig. 2b, d and Extended Data Fig. 2a). The interaction between the epi-DNA recognition loop and 5caC probably causes this positional shift (Fig. 2b–d), which disrupts the proper alignment between Rpb1 L1081 and the substrate, as well as the correct positioning of the 3′-RNA terminus and the substrate that is crucial for full closure of the trigger loop and effective GTP addition[17,18]. The nucleobase of the substrate now misaligns with Rpb1 T831 in the bridge helix (Fig. 2a), leading to a partially open conformation of the trigger loop (Extended Data Fig. 2b).

To determine further whether the specific hydrogen-bonding interaction between the Pol II epi-DNA recognition loop (Rpb2 Q531) and 5caC (Fig. 2b, c) causes a reduction in GTP addition efficiency, we purified two yeast Pol II point mutants (Rpb2 Q531H and Q531A) and measured GTP incorporation on the 5caC template in comparison with wild-type Pol II. The Pol II Q531A mutation abolishes the specific hydrogen bonds between the side chain of residue 531 and

[1]Skaggs School of Pharmacy and Pharmaceutical Sciences, The University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. [2]Department of Cellular and Molecular Medicine, School of Medicine, The University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. [3]Department of Chemistry, Department of Biochemistry and Molecular Biology, and Institute for Biophysical Dynamics, Howard Hughes Medical Institute, The University of Chicago, Chicago, Illinois 60637, USA.
*These authors contributed equally to this work.

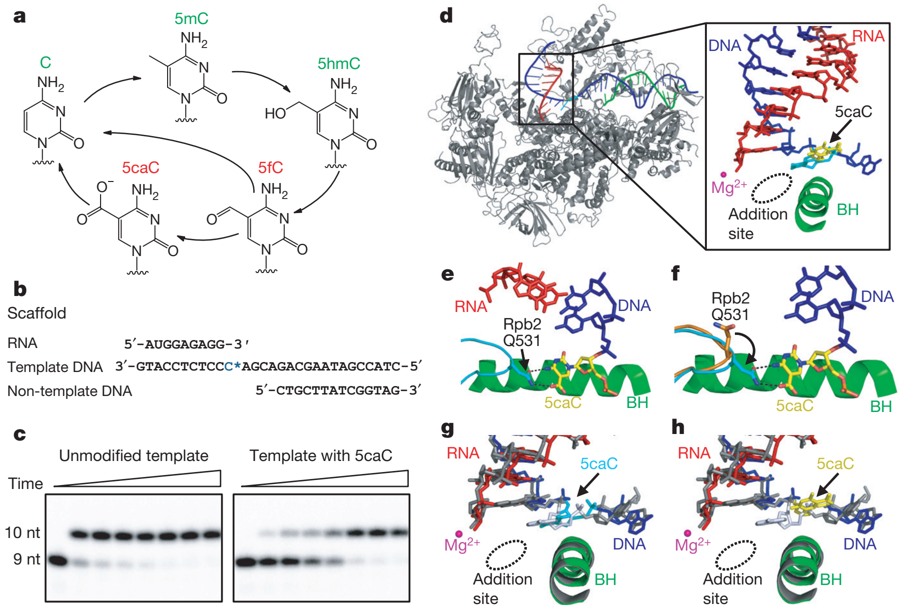


Figure 1 | Pol II directly recognizes 5caC during transcription. **a**, Epigenetic modification cycle of cytosine. **b**, The RNA/DNA scaffold used in both structural and biochemical analysis. C* indicates 5caC residue. **c**, Impeded Pol II elongation on the 5caC-containing template relative to the unmodified C template. Time points are 0, 5 s, 15 s, 30 s, 1 min, 5 min, 20 min, and 1 h (left to right). nt, nucleotides. **d**, The overall Pol II EC structure containing a site-specific 5caC (EC-I). Colour-coded are template DNA (blue), non-template DNA (green) and RNA (red). The two 5caC conformers are highlighted in yellow and cyan, respectively. Part of the bridge helix (BH; Rpb1 822–840) is highlighted in green and the rest of the Pol II subunits are in grey (Rpb2 is omitted). The addition site is represented by a dotted oval. **e**, The midway 5caC interacts with the Rpb2 Q531 residue via hydrogen bonds (black dotted lines). The epi-DNA recognition loop (fork loop 3; Rpb2 521–541) is shown in cyan. **f**, The Q531 side chain rotates 90° to form hydrogen bonds with 5caC. Pol II EC-I is superimposed with the Pol II EC containing an unmodified DNA template in the post-translocation state (PDB accession 1SFO). The fork loop 3 region of Pol II EC (PDB accession 1SFO) is shown in orange. **g**, **h**, Comparison of two 5caC conformers (cyan or yellow) with the corresponding canonical template nucleotide (blue-white).

the 5-carboxyl moiety of 5caC, thus alleviating the negative impact of 5caC on Pol II transcription. In contrast, the Q531H mutant should behave similarly to wild-type Pol II, as the His residue has the capability to form a hydrogen bond with 5caC. Indeed, we observed that the Q531A mutant leads to a 2.6-fold increase in GTP incorporation specificity (catalytic rate constant (k_{pol})/apparent dissociation constant ($K_{\mathrm{d,app}}$)) (P value <0.0001, unpaired, two-sided t-test) with a faster k_{pol} and a tighter $K_{\mathrm{d,app}}$, whereas the Q531H mutant behaves like wild-type Pol II (Fig. 2e–g and Extended Data Fig. 3). Consistently, abolishing the specific hydrogen-bonding interaction with the Pol II epi-DNA recognition loop by removal of the 5-carboxyl moiety of 5caC (replacement of 5caC to unmodified C template) also leads to a 4.2-fold increase in GTP incorporation specificity[14]. These data demonstrate the functional impact of 5caC on both the rate and specificity of GTP incorporation via Q531 in the Pol II epi-DNA recognition loop.

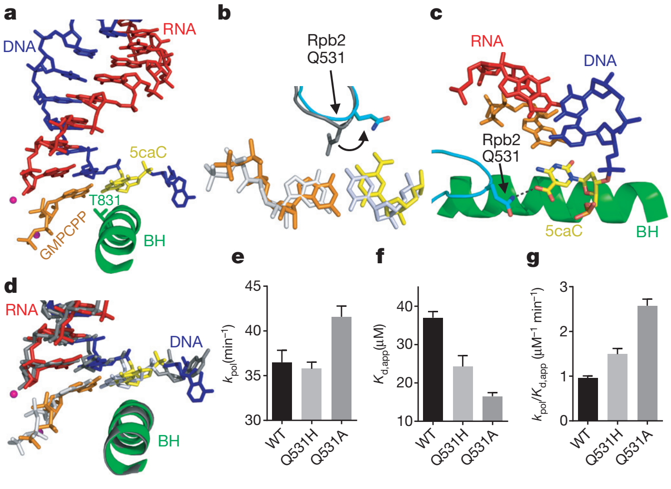


Figure 2 | Interaction between 5caC and epi-DNA recognition loop compromises GTP incorporation. **a**, The Pol II EC structure containing a matched GMPCPP opposite the 5caC site (EC-II). The colour codes are the same as Fig. 1 except for 5caC (yellow) and GMPCPP (orange). **b–d**, The GMPCPP:5caC base pair is shifted towards the downstream main channel from the canonical GMPCPP:dC position (PDB accession 2E2J). The side chain of Rpb2 Q531 rotates 100° to interact with 5caC (**b**, **c**). **e–g**, Comparison of catalytic rate constants (k_{pol}) (**e**), substrate dissociation constants $K_{\mathrm{d,app}}$ (**f**) and specificity constants ($k_{\mathrm{pol}}/K_{\mathrm{d,app}}$) (**g**) of GTP incorporation opposite the 5caC template by wild-type (WT), Q531H and Q531A Pol II, respectively. The mean values are presented and error bars are standard deviations derived from three independent experiments.

The same epi-DNA recognition loop is equally capable of forming similar interactions with the 5-carbonyl group of 5fC, but not with 5mC or unmodified C (Extended Data Fig. 4). We also modelled the structure of 5hmC (Protein Data Bank (PDB) accession 4R2C) within the Pol II EC. Interestingly, the 5-hydroxyl group appears to adopt a different orientation away from the epi-DNA recognition loop (Extended Data Fig. 4c), consistent with less Pol II retardation on a 5hmC template relative to a 5fC/5caC template[14]. Intriguingly, a recent study revealed that base J (β-D-glucosyl-hydroxymethyluracil), a major groove DNA modification, prevents transcriptional read-through in *Leishmania*[19]. We observed a striking similarity between 5fC/5caC and base J on slowing down or stalling Pol II transcription, suggesting a likely universal mechanism for Pol II to sense distinct DNA modifications via the epi-DNA recognition loop[15].

The critical Gln residue (Q531 in yeast Rpb2) is conserved among several fungal species containing active TET/JBP enzymes and oxi-mCs, such as *Agaricomycetes* and *Pucciniomycete*, and is substituted by the functionally equivalent His residue in mammals (Extended Data Fig. 5a)[20], suggesting that similar interactions between 5caC and the Pol II epi-DNA recognition loop are probably maintained from fungi to humans (Extended Data Fig. 5b–d). Indeed, we observed impeded elongation by human Pol II in HeLa nuclear extracts (Extended Data Fig. 6) and with purified rat Pol II on the 5caC-containing template relative to the unmodified C template[14]. The conservation of this critical Gln/His residue in eukaryotes coincides with the existence of 5fC/5caC modifications. In contrast, bacteria and archaea RNA polymerases carry Ala or Pro at the corresponding position in the βD loop II region (Extended Data Fig. 5a)[21], and consistently, we found that 5caC has no observable effect on *Escherichia coli* RNA polymerase transcription elongation *in vitro* (Extended Data Fig. 7, bottom). It is interesting to note that glycosylated cytosine derivatives are also present in some phage and bacterial DNA genomes, pointing to future investigations to understand how these modifications (bulkier than 5fC/5caC) may be recognized during transcription.

Our structural studies also shed light on the canonical Pol II translocation process by revealing two new translocation intermediate positions of the 5caC template before and after GTP binding. The first translocation intermediate position of 5caC sits above the bridge helix in the absence of GTP. Upon GTP binding, the 5caC template is shifted to a new translocation intermediate position allowing the formation of a base pair with incoming GTP. The translocation intermediate states are similar to the translocation intermediate states on an unmodified DNA template recently suggested by molecular simulation[22]. Our ability to capture the crystal structures of these Pol II translocation intermediates suggests that the specific interactions between the Pol II recognition loop (Q531) and the 5-carboxyl group of 5caC stabilize the translocation intermediates that are otherwise too transient to be captured on unmodified DNA template.

Aligning the structures of 5caC-paused Pol II EC with bulky DNA lesion-arrested or α-amanitin-arrested Pol II EC[18,23,24] reveals additional insights into Pol II pausing and arrest. Notably, 5caC, CPD, pyriplatin-dG and the i+1 transition template base in α-amanitin-arrested Pol II EC are all accommodated above the bridge helix (Fig. 3a–c), even though their exact locations, orientations and interactions with Pol II greatly differ (see Methods). A similar (but not identical) 'above-the-bridge-helix' translocation intermediate has also been recently observed in an elemental paused *E. coli* RNA polymerase structure (ePEC) with a kinked bridge helix to occlude the canonical i+1 template position[25]. Taken together, we propose that these observations point to a common translocation checkpoint that serves as a rate-limiting step for the transition of the DNA template nucleobase to cross over the bridge helix and subsequently insert into the canonical i+1 site to guide RNA synthesis. While DNA lesions have been proposed to interfere with Pol II elongation via steric hindrance[26,27], our current data suggest that Pol II can also directly sense epigenetically modified DNA (5caC/5fC) through specific hydrogen-bonding interactions.

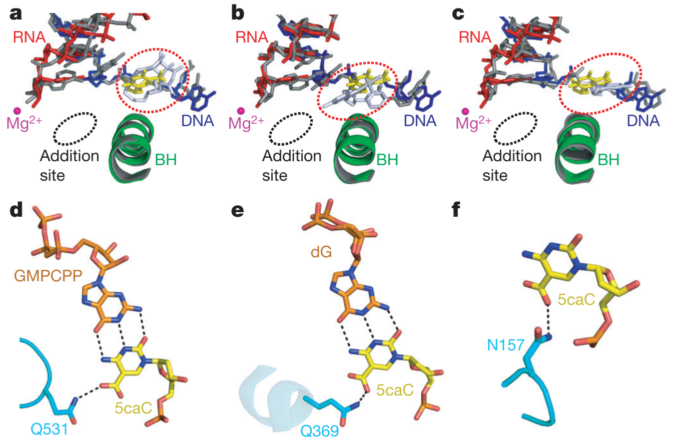


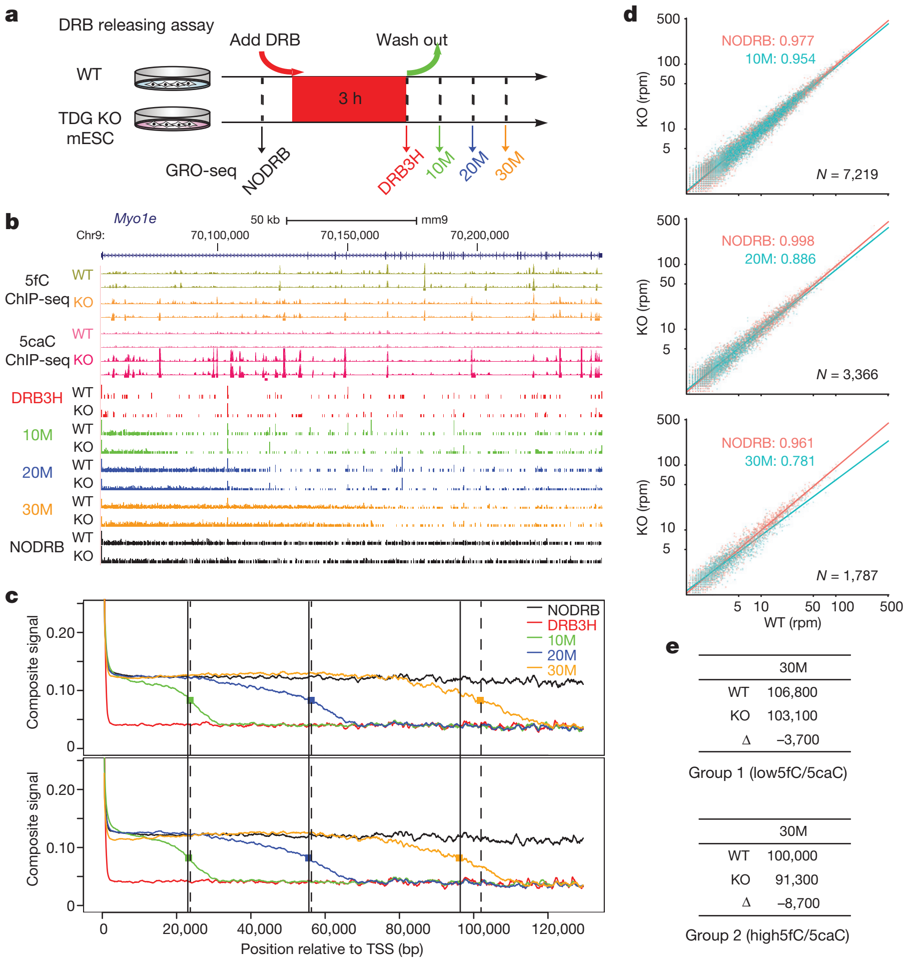
Figure 3 | Similar 'above-the-bridge-helix' translocation intermediates captured in pausing/arrested Pol II ECs and a common 5caC-recognition mode shared by a variety of 5caC-recognition proteins. a–c, Superimposition of 5caC-paused Pol II EC with CPD-lesion-arrested EC (PDB accession 4A93) (**a**), pyriplatin-lesion-arrested EC (PDB accession 3M4O) (**b**), and α-amanitin-arrested EC (PDB accession 2VUM) (**c**). The similar above-the-bridge-helix translocation intermediates region for accommodation of the i+1 5caC (yellow) and DNA lesion (or translocation intermediate captured by α-amanitin) (blue-white) is highlighted by a red-dotted oval. The damage-arrested or α-amanitin-arrested Pol II ECs are shown in grey. BH, bridge helix. **d–f**, The conserved interactions and residues involved in 5caC recognition by Pol II (Rpb2 Q531) (**d**), human WT1 (Q369; PDB accession 4R2R) (**e**), and human TDG (N157; PDB accession 3UO7) (**f**).

We further noticed a remarkable mechanistic similarity in 5caC recognition by several unrelated family proteins. For example, residue Q369 in human Wilms tumor protein 1 (WT1) (PDB accession 4R2R)[13] and residue N157 in human TDG (PDB accession 3UO7)[28] are both functionally equivalent to Q/H531 residues in Pol II in recognizing the 5caC carboxyl group via specific hydrogen bonds (Fig. 3d–f). We thus speculate that 5caC could be a potential epigenetic mark for recognition by a variety of 'protein readers' (including Pol II itself) via specific hydrogen-bonding interactions with its 5-carboxyl moiety.

To determine further the functional consequences of oxi-mCs on Pol II transcription elongation in mammalian cells, we measured the *in vivo* transcription elongation rate on a pair of isogenic mouse embryonic stem (ES) cells ($Tdg^{fl/fl}$ wild-type mouse ES cells and $Tdg^{-/-}$ knockout mouse ES cells derived from conditional TDG-knockout mice) by global nuclear run-on coupled with deep sequencing (GRO-seq) (Fig. 4a). Previous studies showed that, relative to wild type, TDG knockout led to a substantial increase of global 5fC/5caC levels[7,8]. The GRO-seq experiments allowed us to measure the front edge of waves of nascent transcripts at different time points to deduce the rate of Pol II transcription elongation.

We observed retarded Pol II elongation in TDG-knockout ES cells relative to wild-type ES cells after the functional impact was sufficiently accumulated, as exemplified on the long *Myo1e* gene (Fig. 4b). Further metagene analysis of the middle points of wild-type and TDG-knockout mouse ES cells at different time points revealed a clear reduction of Pol II elongation in TDG-knockout relative to wild-type ES cells after 30 min of synchronized transcription, although the differences at earlier times were not evident (Fig. 4c). We next analysed the GRO-seq read density gene by gene ±10 kb around individual middle points followed by linear regression to determine the slope (Fig. 4d). We observed progressive slowing down of Pol II in TDG-knockout ES cells relative to wild-type cells, as indicated by decreasing slopes, and the read density ratio (TDG knockout/wild type) at 30 min was significantly smaller relative to control (P value $= 1.52 \times 10^{-10}$ from one-sided Kolmogorov–Smirnov test) (Fig. 4d). Finally, to determine the dosage-dependent effect of 5fC/5caC on Pol II transcription elongation, we focused on the data at 30 min and segregated genes into two groups according to increased levels of 5fC/5caC in response to TDG knockout and compared the middle points at individual assay points. The data indicate a correlation between increased 5fC/5caC and a reduced transcription elongation rate among genes in group 2 (high 5fC/5caC level) relative to group 1 (low 5fC/5caC level) (Fig. 4e). Together, these global data demonstrate retarded Pol II elongation by enhanced 5fC/5caC levels in the gene body. The combination of *in vitro* and *in vivo* data strongly indicates a direct impact of 5fC/5caC on Pol II elongation on the DNA template.

We present structural and biochemical evidence to suggest that Pol II has the ability to sense the DNA oxidative methylation state directly through its conserved epi-DNA recognition loop, and that it transiently slows down at oxi-mC (5fC/5caC) sites during transcription. Since 5fC/5caC are not distributed evenly across the genome and show considerable variation between cell types, it is conceivable that these pausing effects may add a new layer of fine-tuned regulation of Pol II transcription elongation dynamics. For example, compared with transcription of short genes, the cumulative consequences of pausing effects at 5fC/5caC sites probably have much more profound regulatory impacts on transcribing some long genes that are preferentially expressed in the brain and have crucial roles in neuronal integrity[29].



	30M
WT	106,800
KO	103,100
Δ	−3,700

Group 1 (low5fC/5caC)

	30M
WT	100,000
KO	91,300
Δ	−8,700

Group 2 (high5fC/5caC)

Figure 4 | Impact of 5fC/5caC on Pol II transcription elongation in mouse ES cells. a, Scheme of the DRB releasing assay. Wild-type (WT) and TDG-knockout (KO) mouse ES cells (mESC) were treated with DRB followed by washing out DRB to allow transcription for 10, 20 or 30 min (10M, 20M and 30M, respectively). No DRB treatment (NODRB) or 3 h DRB treatment (DRB3H) were performed as controls. All experiments were performed in duplicate and reproducibility was evident in all pairwise comparisons (Extended Data Fig. 8). **b**, The GRO-seq data on the representative *Myo1e* gene. Elevated 5fC/5caC levels in TDG-knockout mouse ES cells are derived from published chromatin immunoprecipitation followed by sequencing (ChIP-seq) data in duplicate[8]. Chr, chromosome. **c**, Comparative metagene analysis of GRO-seq signals between wild-type (top) and knockout mouse ES cells (bottom). Dashed and non-dashed lines show the middle points of the ensemble transcription waves in wild-type and knockout mouse ES cells, respectively. TSS, transcription start site. **d**, Pairwise comparisons of the GRO-seq density (reads per million (rpm)) of individual genes in the ±10 kb window around different middle points between wild-type (*x*-axis) and knockout cells (*y*-axis) in **c** (10M, 20M, 30M in cyan) with the NODRB data (red) as control. The coefficients are the slopes of the lines from linear regression on the scattered points. The *P* values were calculated based on one-sided Kolmogorov–Smirnov test of comparing read density ratio (knockout/wild type) at 30 min. *N*, number of genes. **e**, Correlation between increased 5fC/5caC levels and retarded transcription elongation. Genes were divided into two groups according to increased 5fC/5caC levels in the gene bodies (low in group 1 and high in group 2). The numbers correspond to the middle point positions (bp) of the ensemble transcription waves relative to the TSS in wild-type versus knockout mouse ES cells.

The transient Pol II pausing at 5fC/5caC sites may also provide signals for the recruitment of various transcription elongation factors, chromatin remodelling complexes, messenger RNA processing machineries, and TDG and the base excision repair machineries to the oxi-mC sites to induce additional functional consequences. On the basis of the similarity between direct Pol II recognition of 5caC and the role of Pol II in sensing bulky DNA lesions in transcription-coupled nucleotide excision repair, we propose that Pol II may act as a direct sensor for a variety of DNA modification and damage events to instruct distinct downstream pathways[26,27,30].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 June 2014; accepted 20 April 2015.

Published online 29 June 2015.

1. Pastor, W. A., Aravind, L. & Rao, A. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nature Rev. Mol. Cell Biol.* **14,** 341–356 (2013).
2. Wu, H. & Zhang, Y. Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell* **156,** 45–68 (2014).
3. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324,** 930–935 (2009).
4. Pfaffeneder, T. *et al.* The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem. Int. Ed. Engl.* **50,** 7008–7012 (2011).
5. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333,** 1300–1303 (2011).
6. He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333,** 1303–1307 (2011).
7. Song, C. X. *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153,** 678–691 (2013).
8. Shen, L. *et al.* Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153,** 692–706 (2013).
9. Klose, R. J. & Bird, A. P. Genomic DNA methylation: the mark and its mediators. *Trends Biochem. Sci.* **31,** 89–97 (2006).
10. Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38,** 23–38 (2013).
11. Spruijt, C. G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152,** 1146–1159 (2013).
12. Iurlaro, M. *et al.* A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol.* **14,** R119 (2013).

13. Hashimoto, H. *et al.* Wilms tumor protein recognizes 5-carboxylcytosine within a specific DNA sequence. *Genes Dev.* **28,** 2304–2313 (2014).
14. Kellinger, M. W. *et al.* 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. *Nature Struct. Mol. Biol.* **19,** 831–833 (2012).
15. Huang, Y. & Rao, A. New functions for DNA modifications by TET-JBP. *Nature Struct. Mol. Biol.* **19,** 1061–1064 (2012).
16. Cramer, P., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: RNA polymerase II at 2.8 Ångstrom resolution. *Science* **292,** 1863–1876 (2001).
17. Wang, D., Bushnell, D. A., Westover, K. D., Kaplan, C. D. & Kornberg, R. D. Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. *Cell* **127,** 941–954 (2006).
18. Brueckner, F. & Cramer, P. Structural basis of transcription inhibition by α-amanitin and implications for RNA polymerase II translocation. *Nature Struct. Mol. Biol.* **15,** 811–818 (2008).
19. van Luenen, H. G. *et al.* Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in *Leishmania*. *Cell* **150,** 909–921 (2012).
20. Iyer, L. M. *et al.* Lineage-specific expansions of TET/JBP genes and a new class of DNA transposons shape fungal genomic and epigenetic landscapes. *Proc. Natl Acad. Sci. USA* **111,** 1676–1683 (2014).
21. Korzheva, N. *et al.* A structural model of transcription elongation. *Science* **289,** 619–625 (2000).
22. Silva, D. A. *et al.* Millisecond dynamics of RNA polymerase II translocation at atomic resolution. *Proc. Natl Acad. Sci. USA* **111,** 7665–7670 (2014).
23. Wang, D., Zhu, G. Y., Huang, X. H. & Lippard, S. J. X-ray structure and mechanism of RNA polymerase II stalled at an antineoplastic monofunctional platinum-DNA adduct. *Proc. Natl Acad. Sci. USA* **107,** 9584–9589 (2010).
24. Walmacq, C. *et al.* Mechanism of translesion transcription by RNA polymerase II and its role in cellular resistance to DNA damage. *Mol. Cell* **46,** 18–29 (2012).
25. Weixlbaumer, A., Leon, K., Landick, R. & Darst, S. A. Structural basis of transcriptional pausing in bacteria. *Cell* **152,** 431–441 (2013).
26. Lindsey-Boltz, L. A. & Sancar, A. RNA polymerase: the most specific damage recognition protein in cellular responses to DNA damage? *Proc. Natl Acad. Sci. USA* **104,** 13213–13214 (2007).
27. Hanawalt, P. C. & Spivak, G. Transcription-coupled DNA repair: two decades of progress and surprises. *Nature Rev. Mol. Cell Biol.* **9,** 958–970 (2008).
28. Zhang, L. *et al.* Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nature Chem. Biol.* **8,** 328–330 (2012).
29. Polymenidou, M. *et al.* Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nature Neurosci.* **14,** 459–468 (2011).
30. Sarker, A. H. *et al.* Recognition of RNA polymerase II and transcription bubbles by XPG, CSB, and TFIIH: insights for transcription-coupled repair and Cockayne syndrome. *Mol. Cell* **20,** 187–198 (2005).

Acknowledgements D.W. acknowledges the National Institutes of Health (NIH) (GM102362), a Kimmel Scholars award from the Sidney Kimmel Foundation for Cancer Research, and start-up funds from the Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego. This work was also supported by NIH grant HG006827 and the Howard Hughes Medical Institute to C.H., and NIH grants GM052872 and HG004659 to X.-D. F. We are grateful to C. Kaplan for providing *Saccharomyces cerevisiae* Pol II Rpb2 Q531H and Q531A mutant strains.

Author Contributions D.W. conceived the original idea and, together with X.-D.F., designed the experiments. X.L. carried out synthesis of DNA templates. J.C., L.W. and D.W. purified Pol II. L.W. and D.W. performed crystallization, data collection and structural refinement. L.X. performed the *in vitro* transcription assay. Y.Z., R.X., L.C. and H.L. performed the *in vivo* GRO-seq assay. L.W., Y.Z., L.X., R.X., X.L., J.C., C.H., X.-D.F. and D.W. wrote the paper.

Author Information GRO-seq data have been deposited in the Gene Expression Omnibus database under accession GSE64748. Atomic coordinates and structure factors for the reported crystal structures have been deposited in the Protein Data Bank under accessions 4Y52 and 4Y7N for EC-I and EC-II, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.W. (dongwang@ucsd.edu) or X.-D.F. (xdfu@ucsd.edu).

METHODS

Preparation of Pol II ECs. *Saccharomyces cerevisiae* Pol II was purified as previously described[17]. PAGE-purified RNA oligonucleotides were purchased from Dharmacon, non-template DNA oligonucleotides were obtained from IDT, and template DNA oligonucleotides with 5caC were prepared and purified as previously described[14]. The template DNA, non-template DNA and RNA oligonucleotides were annealed to form the scaffold. To form the Pol II EC, Pol II was mixed with scaffold in the reaction buffer (20 mM Tris (pH 7.5), 40 mM KCl and 5 mM dithiothreitol (DTT)). The final concentrations were 2 μM Pol II, 10 μM template DNA and 20 μM non-template DNA and RNA oligonucleotides. The mixture was incubated at room temperature for 1 h, followed by ultrafiltration to remove excess oligonucleotides. The Pol II elongation complex was crystallized using the hanging drop method and from solutions containing 390 mM $(NH_4)_2HPO_4/NaH_2PO_4$, pH 6.0, 50 mM dioxane, 10 mM DTT and 10.7–11.6% (w/v) PEG6000. Crystals were transferred in a stepwise manner to cryo buffer as previously described[17]. For the Pol II EC with GMPCPP, Pol II EC crystals were soaked with 5–10 mM GMPCPP and 10 mM $MgCl_2$ overnight before harvest.

Data collection and structure determination of 5caC-paused Pol II ECs. Diffraction data were collected on beamlines 8.2.1 and 5.0.2 at the Advanced Light Source, Lawrence Berkeley National Laboratory. Data were processed in DENZO and SCALEPACK (HKL2000)[31]. Model building was performed with the program Coot[32], and refinement was done with REFMAC5 with TLS (CCP4i) or PHENIX (Extended Data Table 1)[33]. Electron density maps are shown in Extended Data Fig. 1. EC-I refers to the Pol II EC crystal structure that contains a site-specific 5caC at the i+1 site in the absence of GTP binding. EC-II refers to the Pol II EC crystal structure that contains a site-specific 5caC at the i+1 site in the presence of GMPCPP. Ramachandran plot of EC-I showed 85.57%, 11.70% and 2.73% of EC-I residues are in preferred, allowed and disallowed regions, respectively. For EC-II, 86.00%, 11.22%, and 2.78% of residues are in above regions, respectively. All structural models in the figures were superimposed with the bridge helix region (Rpb1 822–840) near the active site using Coot[32] and PyMOL[34].

Pol II purification for *in vitro* transcription assay. *S. cerevisiae* Pol II and mutants were purified essentially as previously described[17]. Briefly, Pol II (with recombinant protein A tag at Rpb3 subunit) was first affinity-purified by IgG column. The Pol II elution from IgG column was further purified using HiTrap Heparin and Mono Q (GE Healthcare). The final pure Pol II (Extended Data Fig. 3d) was ready for future *in vitro* transcription experiments.

***In vitro* transcription assays.** The *S. cerevisiae* Pol II ECs for transcription assays were assembled using established methods[16]. Briefly, an aliquot of 5′-^{32}P-labelled RNA was annealed with a 1.5-fold amount of template DNA and a 2-fold amount of non-template DNA to form RNA/DNA scaffold in elongation buffer (20 mM Tris-HCl, pH 7.5, 40 mM KCl and 5 mM $MgCl_2$). An aliquot of annealed scaffold of RNA/DNA was then incubated with a fourfold excess amount of Pol II at room temperature for 10 min to ensure the formation of Pol II EC. The *in vitro* transcription started when the Pol II EC was mixed with equal volumes of GTP solution. The final concentrations were 25 nM scaffold, 100 nM Pol II and 1 μM GTP in the elongation buffer. Reactions were quenched at various time points by the addition of one volume of 0.5 M EDTA (pH 8.0). (Time points are 0, 5 s, 15 s, 30 s, 1 min, 5 min, 20 min, and 1 h). The quenched products were analysed by denaturing PAGE and visualized using a storage phosphor screen and Pharos FX imager (Bio-Rad). The *in vitro* transcription assay of *E. coli* RNA polymerase (RNAP, New England Biolabs (NEB)) was performed using the same procedure as *S. cerevisiae* RNA Pol II transcription.

For the transcription of human Pol II in the nuclear extract of HeLa cells (Life Technologies), the excess annealed scaffold was incubated with nuclear extract of HeLa cells for 5 min before the addition of α-^{32}P-GTP. The final concentrations were 1 μM scaffold, 1 μM α-32P-GTP (0.2 μCi μl^{-1}) and 3 mg ml^{-1} protein of nuclear extract. Reactions were then quenched at various time points by the addition of one volume of 0.5 M EDTA (pH 8.0). The quenched products were analysed by denaturing PAGE and visualized using a storage phosphor screen and Pharos FX imager (Bio-Rad). All transcription assays described earlier were performed independently in triplicates.

***In vitro* RNA pol II transcription kinetic assay and analysis.** The assay was carried out as previously described[14]. Briefly, nucleotide incorporation assays were conducted by pre-incubating 50 nM annealed scaffold containing a site-specific 5caC modification at the template with 200 nM purified Pol II (wild type, Q531H and Q531A) for 10 min in elongation buffer at room temperature. The Pol II EC was then mixed with an equal volume of solution containing 40 mM KCl, 20 mM Tris-HCl (pH 7.5), 10 mM DTT, 10 mM $MgCl_2$ and twofold concentrations of various nucleotides. Final reaction concentrations after mixing were 25 nM scaffold, 100 nM Pol II, 5 mM $MgCl_2$ and various nucleotide concentrations in elongation buffer. Reactions were quenched at various times by addition of one volume of 0.5 M EDTA (pH 8.0) and analysed by denatured PAGE.

Nonlinear-regression data fitting was performed using Prism 6. The time dependence of product formation was fit to a one-phase association equation (product $= Ae^{(-k\text{obs } t)} + C$) to determine the observed rate (k_{obs}). The substrate concentration dependence was fit to a hyperbolic equation ($k_{\text{obs}} = k_{\text{pol}}$ [substrate]/($K_{\text{d,app}}$ + [substrate])) to obtain values for the maximum rate of NTP incorporation (k_{pol}) and apparent K_{d} ($K_{\text{d,app}}$) governing NTP binding essentially as described. The specificity constant was determined by $k_{\text{pol}}/K_{\text{d,app}}$.

Cell culture and *in vivo* transcription rate measurement. Wild-type mouse ES cells (*Tdg*$^{fl/fl}$) and TDG-knockout mouse ES cells (*Tdg*$^{-/-}$)[7] were cultured in KnockoutTM DMEM (Life Technologies, catalogue no. 10828-018) supplemented with 15% KnockoutTM Serum Replacement (Life Technologies, catalogue no. 10828-028), 2 mM L-glutamine (Life Technologies, catalogue no. 25030-081), 1× non-essential amino acids (Life Technologies, catalogue no. 11140-050), 1× penicillin–streptomycin (Life Technologies, catalogue no. 15140-122), 0.1 mM 2-mercaptoethanol (Life Technologies, catalogue no. 21985-023), 1,000 U ml^{-1} LIF (Millipore, catalogue no. ESG1106), 3 μM CHIR99021 (Stemgent, catalogue no. 04-0004) and 1 μM PD0325901 (Stemgent, catalogue no. 04-0006). The DRB releasing GRO-seq assays were carried out in mouse ES cells under both wild-type and TDG-knockout conditions. For each time-course assay, there are five samples prepared for GRO-seq: (1) NODRB (without DRB treatment); (2) DRB3H (DRB treatment for 3 h, and this is the 0 time point); (3) 10M (10 min after washing out DRB); (4) 20M (20 min after washing); (5) 30M (30 min after washing). For DRB treatment, we grew cells in a 10 cm plate to 70–80% confluence, treated cells by addition of DRB (Sigma) at a final concentration of 100 mM to the culture medium and incubated for 3 h in the incubator, removed DRB by quick washing cells three times with PBS, then incubated in fresh medium in the incubator to different time points. GRO-seq was implemented as previously described[35,36], and the GRO-seq libraries were subjected to Illumina Hiseq 2000 and 2500 for sequencing.

For each sequencing sample, the sequenced reads were trimmed by removing 3′ adaptor and polyA sequences, and only those longer than 16 bp were used to map the mouse genome (mm9) with Bowtie (version 0.12.7), with parameters "-best-strata -l25 -n2 -k1 -m10"[37]. Only one read was kept for those reads mapped to the same location and strand.

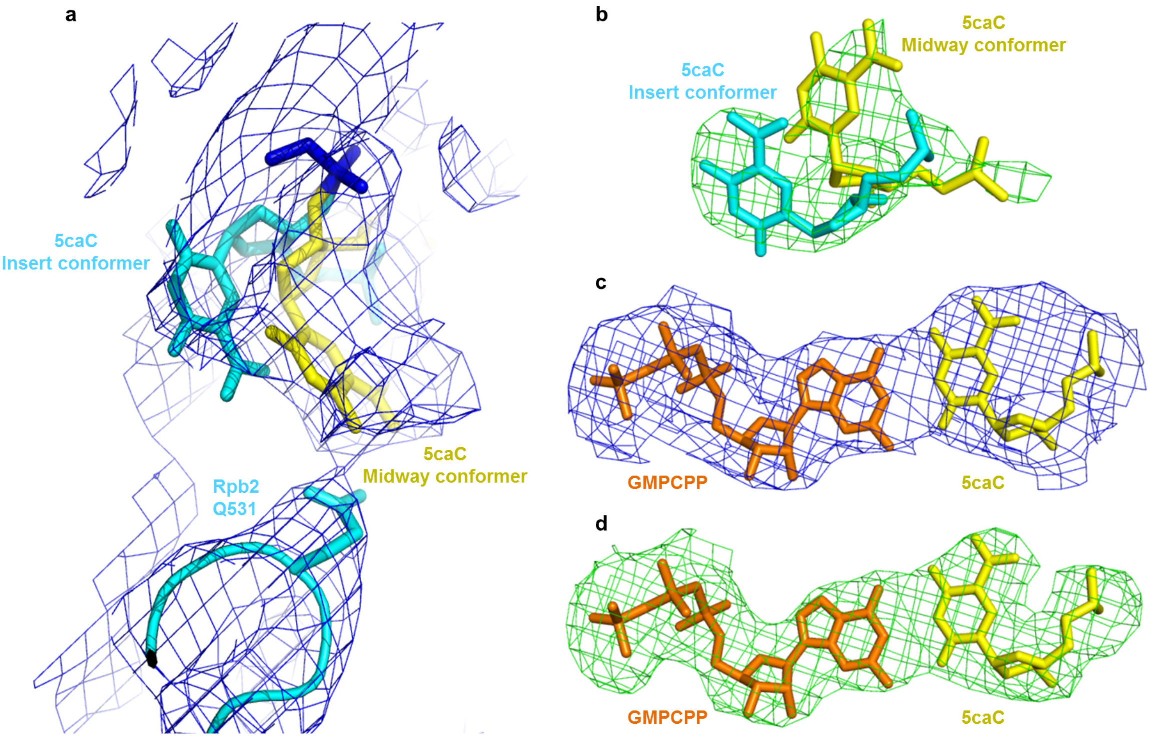
To measure the concordance between replicated samples, we counted the number of the GRO-seq reads in all annotated genes (UCSC refGene) and did pairwise comparisons[38]. Transcripts with the same start and end positions were used once. Having established the data reproducibility, we combined replicated data sets for comparison between biological conditions at different assay points.

To estimate the Pol II elongation rates, we computed the metagene profiles for all assay points. Only the genes with RPKM ≥ 0.5 in the NODRB sample were kept for meta-analysis. The genes were aligned at TSSs, and mapped reads were counted in 100 bp bins across the gene bodies. The counts were normalized to one million total reads per sample, and were averaged for each bin by the number of covering genes and normalized by the relative gene expression in the NODRB sample. The meta-profiles from the normalized counts were smoothed with a 1 kb moving window. The middle point of the ensemble transcription wave at each time point after washing DRB was computed as the position at which the signal reached half of that in the NODRB control sample.

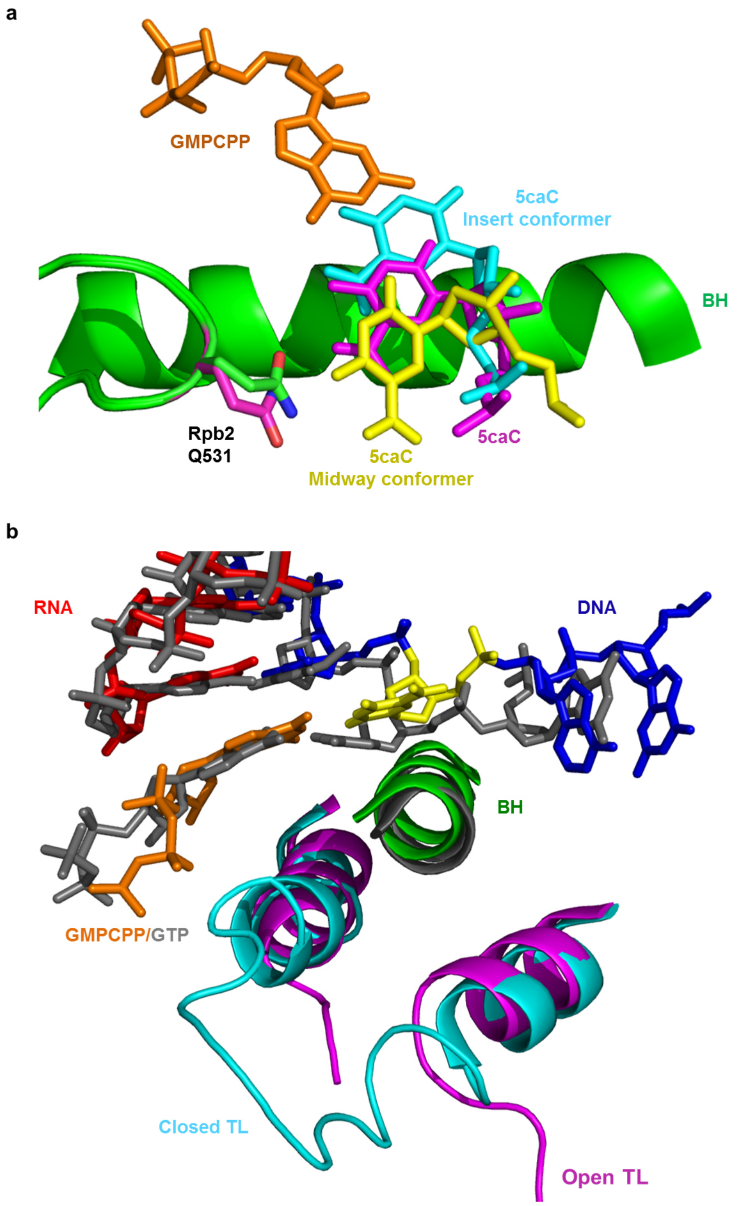
To compare elongation differences at different assay points on individual genes, we calculated the GRO-seq read density in ±10 kb window around the middle points identified in wild-type mouse ES cell cells. At each assay point (10M, 20M, 30M), the counts for each gene in wild-type and TDG-knockout conditions were pairwise compared and linear regressions were fitted to check the trend of change. The samples without DRB treatment were used as control. The changes of 5fC/5caC levels on genes were calculated as the differences of normalized ChIP-seq signals under 5fC/5caC ChIP-seq peaks from knockout to wild type based on the published ChIP-seq data from ref. 8, and the genes were divided into two groups with low and high increased 5fC/5caC levels.

Comparison of 5caC-paused, DNA lesion-arrested, and α-amanitin-arrested Pol II EC. All the structures were aligned by superimposition of the Pol II bridge helix region (residues 822–840 in Rpb1). The i+1 5caC, CPD, pyriplatin-dG and i+1 transition template base in α-amanitin-arrested Pol II EC are all accommodated above the bridge helix, even though their exact locations, orientations and interactions with Pol II greatly differ: α-amanitin appears to capture the Pol II translocation intermediate indirectly by jamming the movement of the Pol II bridge helix and trapping the trigger loop in an inactive conformation, whereas the conformation of CPD and pyriplatin-DNA lesions is largely governed by their covalent crosslink or bulky ligand. In contrast to all of these previous cases, the translocation intermediate of 5caC nucleobase forms a direct interaction with the Pol II epi-DNA recognition loop. Second, the upstream RNA/DNA hybrid adopts essentially the same post-translocation register for all of these paused/arrested Pol II ECs, except that the 3′-RNA/DNA hybrid is substantially tilted for CPD-lesion-arrested Pol II EC. Finally, while 5caC and pyriplatin-DNA lesion can form Watson–Crick base pairs with incoming NTP, thus allowing template-dependent nucleotide addition, the CPD-DNA lesion fails to form such a base pair with the incoming nucleotide, therefore only allowing template-independent ATP incorporation. In contrast to all of these previous cases, the translocation intermediate of the 5caC nucleobase forms a direct interaction with the Pol II epi-DNA recognition loop.

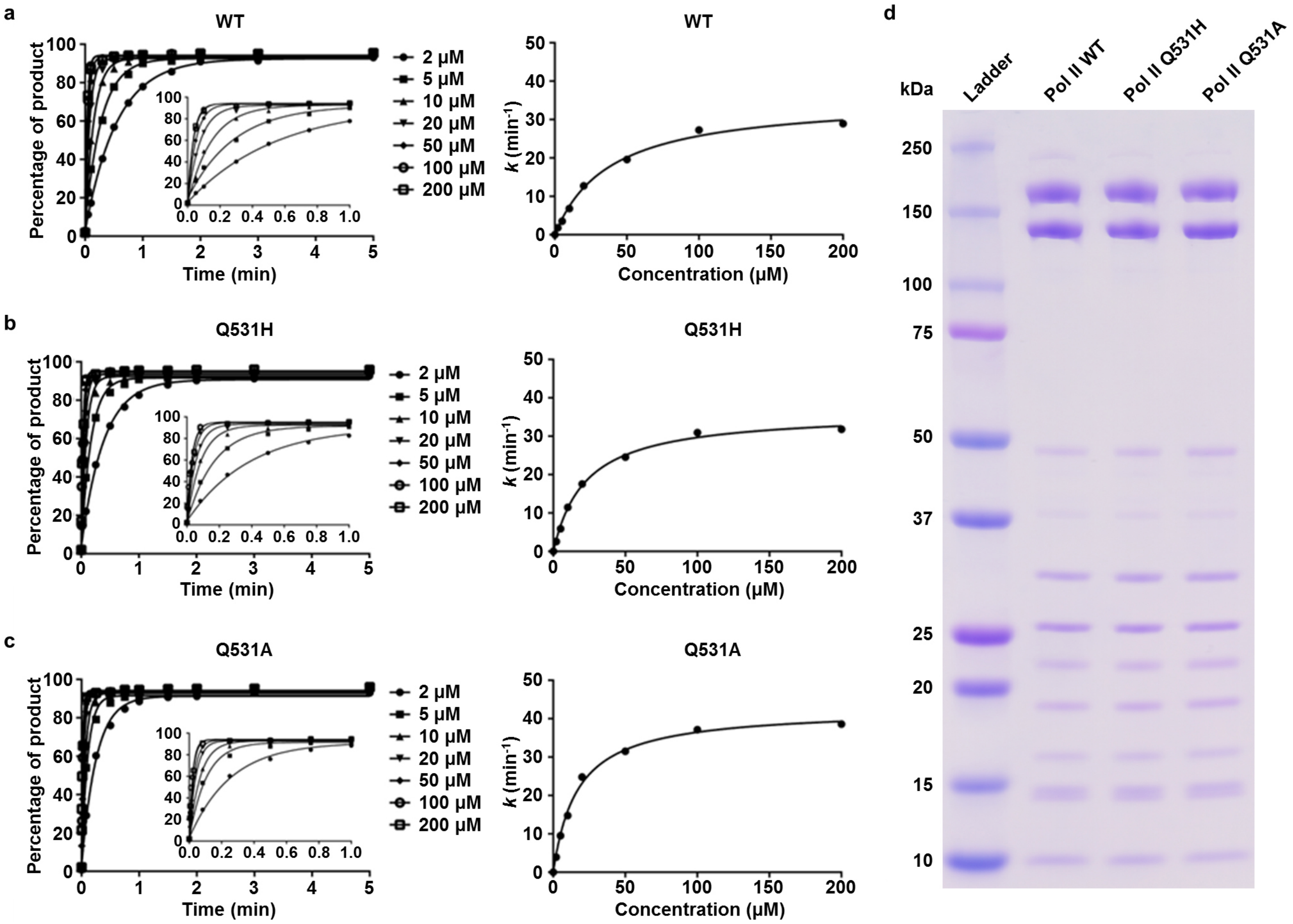
31. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
32. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
33. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
34. DeLano, W. L. *The PyMOL Molecular Graphics System* (DeLano Scientific, 2002).
35. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322,** 1845–1848 (2008).
36. Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474,** 390–394 (2011).
37. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
38. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42,** D764–D770 (2014).



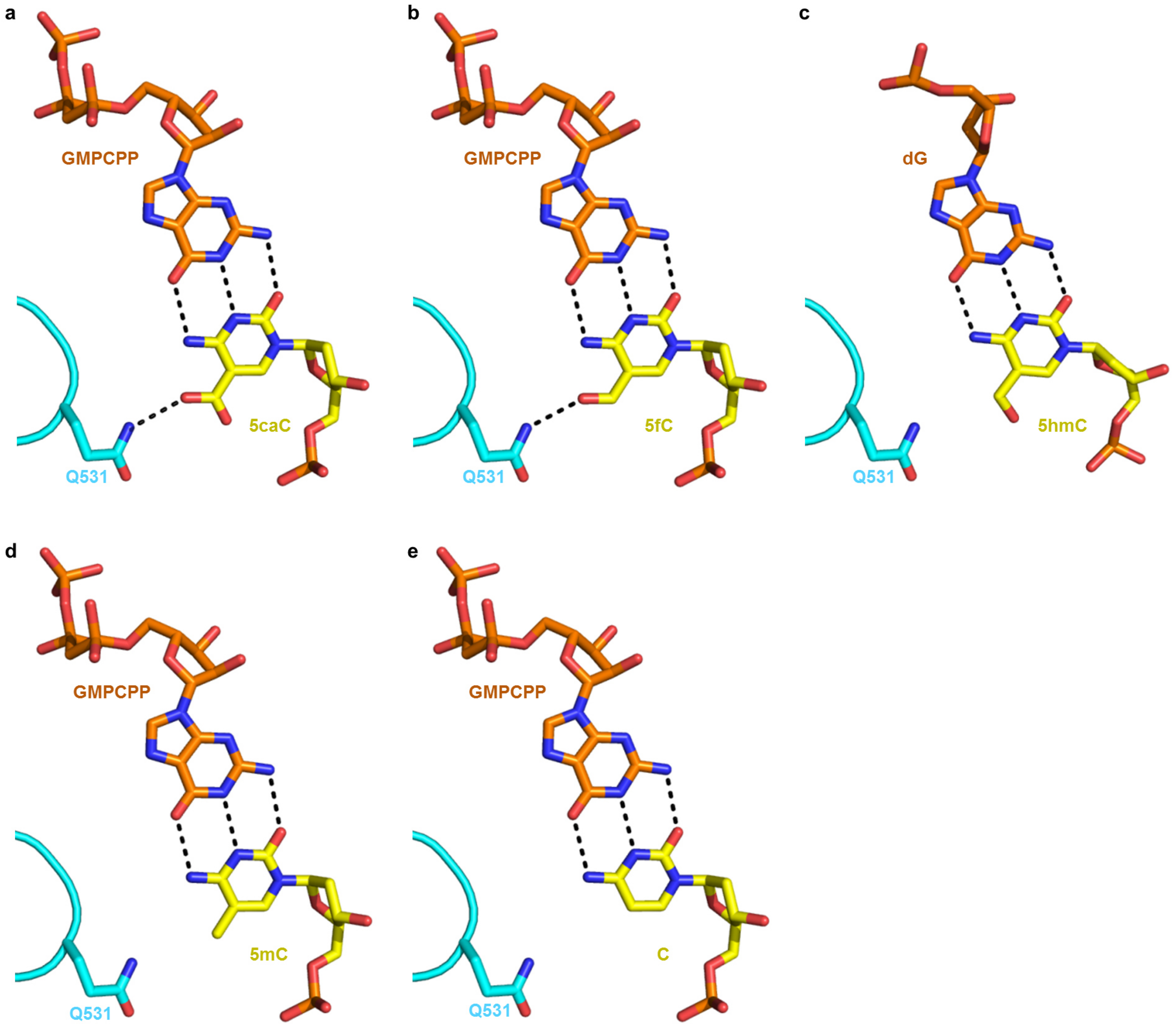
Extended Data Figure 1 | Electron density maps of Pol II EC-I and EC-II. **a**, $2F_o - F_c$ map (blue) of Rpb2 Q531 in epi-DNA recognition loop and the opposite 5caC in Pol II EC-I, contoured at 1.0σ. **b**, $F_o - F_c$ omit map (green) of Pol II EC-I (with 5caC omission), contoured at 3.0σ. **c**, $2F_o - F_c$ map (blue) of GMPCPP paired with 5caC in Pol II EC-II, contoured at 1.0σ. **d**, $F_o - F_c$ omit map (green) of Pol II EC-II (with GMPCPP and 5caC omission), contoured at 3.0σ.



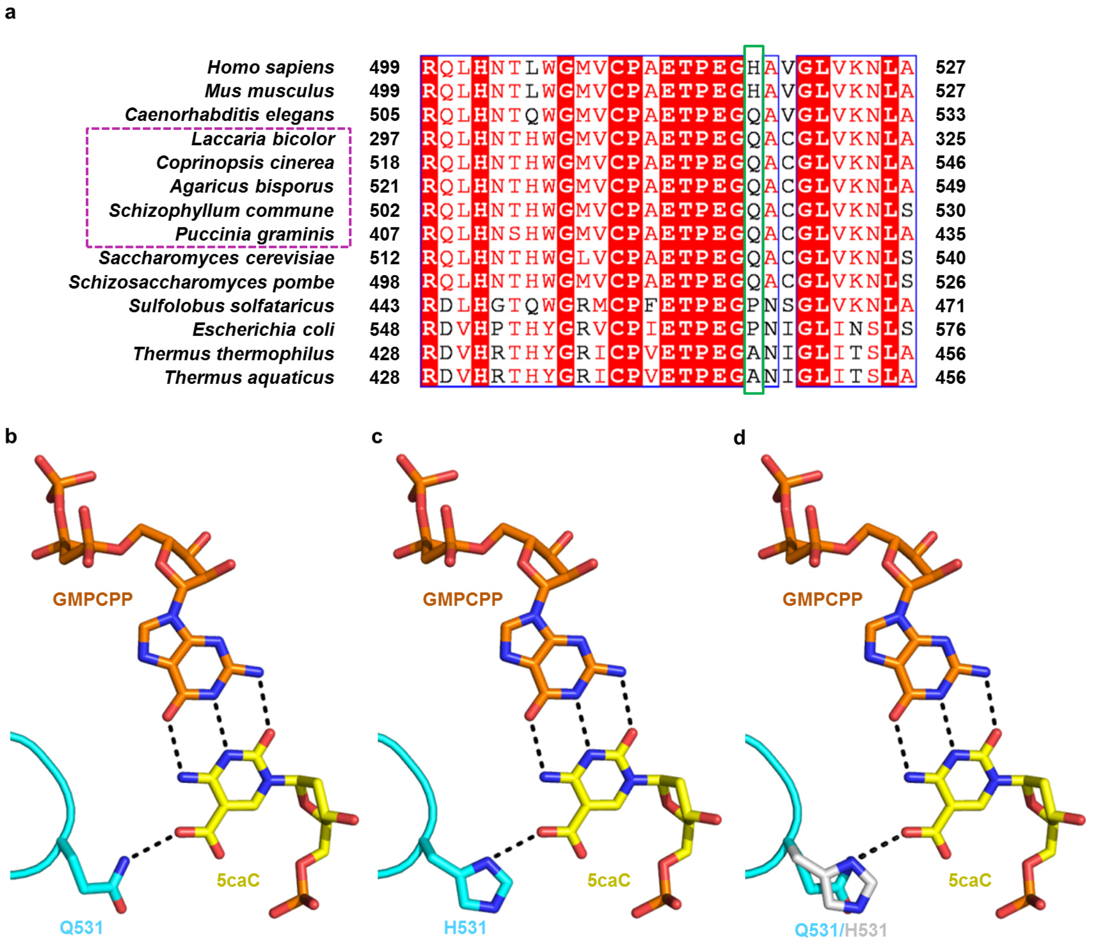
Extended Data Figure 2 | Structural comparison between Pol II EC-I, EC-II and Pol II EC containing unmodified C template and a matched GTP. **a**, Superimposition of Pol II EC-I and EC-II structures. Rpb2 Q531 and 5caC in EC-II are in magenta to differentiate between those counterparts in EC-I. These two structures are aligned using the bridge helix (BH) region (Rpb1 822–840). **b**, Superposition of Pol II EC-II containing 5caC template and GMPCPP with Pol II EC with closed trigger loop (TL; containing unmodified C template and GTP; PDB accession 2E2H). The two structures are aligned using the bridge helix region (Rpb1 822–840).



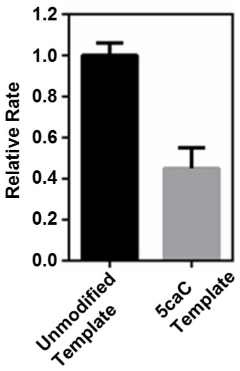
Extended Data Figure 3 | Kinetic study of GTP incorporation opposite 5caC template by purified Pol II proteins. **a**–**c**, Representative kinetic parameter fitting curves from three independent experiments for GTP incorporation opposite 5caC template for Pol II wild type (WT; **a**), Pol II Q531H (**b**) and Pol II Q531A (**c**). **d**, Purified Pol II wild-type, Pol II Q531H and Pol II Q531A proteins used in the *in vitro* transcription experiments.



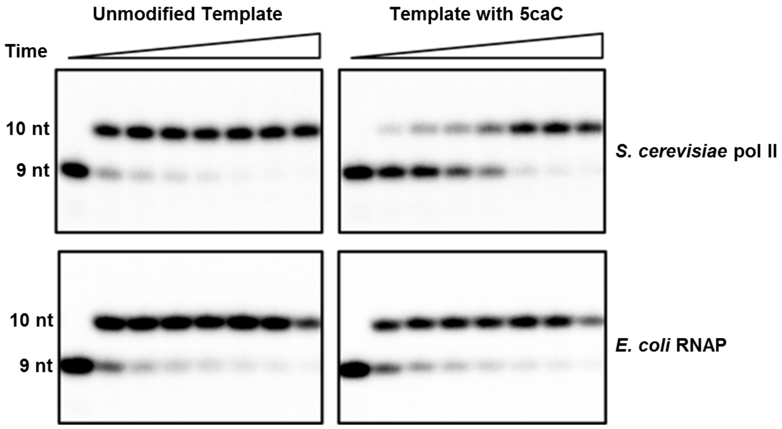
Extended Data Figure 4 | Modelling potentially similar interactions for recognition of 5fC and 5caC templates, but not for 5hmC, 5mC and C templates. **a**, Hydrogen bonds (black dotted lines) between Rpb2 Q531, 5caC and GMPCPP in EC-II. **b**, Model of the interaction between Pol II EC with 5fC template through the same hydrogen-bond interaction network. **c**, Model of Pol II EC with 5hmC template reveals no obvious hydrogen bonding between Q531 and 5hmC. The 5hmC nucleotide structure was based on PDB accession 4R2C. **d**, Model of Pol II EC with 5mC template. **e**, Model of Pol II EC with unmodified C template. The above models were derived from the Pol II EC-II structure.


a
Homo sapiens 499 RQLHNTLWGMVCPAETPEGHAVGLVKNLA 527
Mus musculus 499 RQLHNTLWGMVCPAETPEGHAVGLVKNLA 527
Caenorhabditis elegans 505 RQLHNTQWGMVCPAETPEGQAVGLVKNLA 533
Laccaria bicolor 297 RQLHNTHWGMVCPAETPEGQACGLVKNLA 325
Coprinopsis cinerea 518 RQLHNTHWGMVCPAETPEGQACGLVKNLA 546
Agaricus bisporus 521 RQLHNTHWGMVCPAETPEGQACGLVKNLA 549
Schizophyllum commune 502 RQLHNTHWGMVCPAETPEGQACGLVKNLS 530
Puccinia graminis 407 RQLHNSHWGMVCPAETPEGQACGLVKNLA 435
Saccharomyces cerevisiae 512 RQLHNTHWGLVCPAETPEGQACGLVKNLS 540
Schizosaccharomyces pombe 498 RQLHNTHWGMVCPAETPEGQACGLVKNLS 526
Sulfolobus solfataricus 443 RDLHGTQWGRMCPFETPEGPNSGLVKNLA 471
Escherichia coli 548 RDVHPTHYGRVCPIETPEGPNIGLINSLS 576
Thermus thermophilus 428 RDVHRTHYGRICPVETPEGANIGLITSLA 456
Thermus aquaticus 428 RDVHRTHYGRICPVETPEGANIGLITSLA 456
b
GMPCPP
5caC
Q531
c
GMPCPP
5caC
H531
d
GMPCPP
5caC
Q531/H531

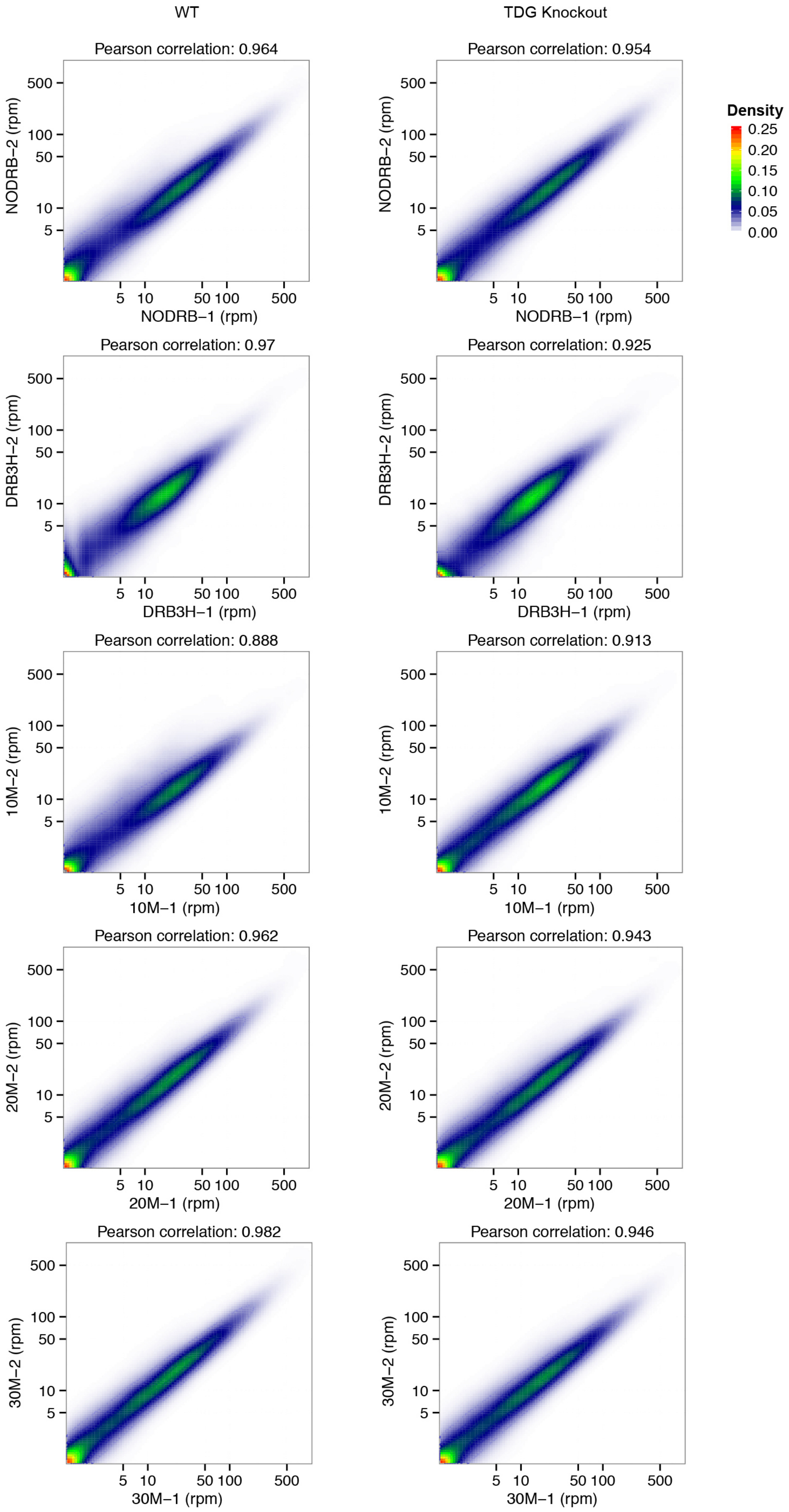
Extended Data Figure 5 | Sequence alignment of Pol II epi-DNA recognition loop across different species. **a**, Pol II epi-DNA recognition loop (Rpb2 521–541) is conserved from fungi to human and strictly conserved among several fungal species, highlighted with magenta dotted rectangle, which contain active TET/JBP enzymes[18]. Key residues in the loop are highlighted in the green box. **b**, Hydrogen bonds (black dotted lines) between yeast Pol II Rpb2 Q531, 5caC and GMPCPP in EC-II. **c**, Model of human Pol II with the functionally equivalent His substitution based on EC-II structure. **d**, Comparison between Q531 and H531 substitution reveals the similar hydrogen-bonding interaction.



Extended Data Figure 6 | Human Pol II slows down at 5caC template in comparison with unmodified template in the context of HeLa nuclear extract. The relative transcription elongation rate is normalized by the transcription elongation rate (k_{obs}) from unmodified template. The relative rates from unmodified template and 5caC template are coloured in black and grey, respectively. The error bars are standard deviations derived from three independent experiments.



Extended Data Figure 7 | Comparison transcription on 5caC template with unmodified template using purified yeast Pol II and *E. coli* RNAP. Top, comparison of yeast Pol II; bottom, comparison of *E. coli* RNAP. Time points are 0, 5 s, 15 s, 30 s, 1 min, 5 min, 20 min, and 1 h (left to right). The top panel is identical to Fig. 1c and is placed here for direct comparison. nt, nucleotides.


WT
TDG Knockout
Density
0.25
0.20
0.15
0.10
0.05
0.00
Pearson correlation: 0.964
NODRB–2 (rpm)
NODRB–1 (rpm)
Pearson correlation: 0.954
NODRB–2 (rpm)
NODRB–1 (rpm)
Pearson correlation: 0.97
DRB3H–2 (rpm)
DRB3H–1 (rpm)
Pearson correlation: 0.925
DRB3H–2 (rpm)
DRB3H–1 (rpm)
Pearson correlation: 0.888
10M–2 (rpm)
10M–1 (rpm)
Pearson correlation: 0.913
10M–2 (rpm)
10M–1 (rpm)
Pearson correlation: 0.962
20M–2 (rpm)
20M–1 (rpm)
Pearson correlation: 0.943
20M–2 (rpm)
20M–1 (rpm)
Pearson correlation: 0.982
30M–2 (rpm)
30M–1 (rpm)
Pearson correlation: 0.946
30M–2 (rpm)
30M–1 (rpm)
500
100
50
10
5
5
10
50
100
500

Extended Data Figure 8 | Correlation between two replicates of GRO-seq data sets at different assay points. GRO-seq replicates (−1 and −2) were pairwise compared gene by gene on the normalized number of reads for wild-type (WT; left) and TDG-knockout (KO; right) samples. The colours show the density of points or genes. The Pearson correlation coefficients were calculated from the points and are shown on the top of each subfigure. rpm, reads per million total reads.

Extended Data Table 1 | Data collection and refinement statistics

	EC-I	EC-II
Data collection		
Space group	C2	C2
Cell dimensions		
a, b, c (Å)	166.7, 221.6, 192.4	168.2, 222.6, 192.8
α, β, γ (°)	90, 100.4, 90	90, 101.6, 90
Resolution (Å)	50-3.5 (3.56-3.5) *	50-3.3 (3.36-3.3)
R_{sym}	0.143 (0.583)	0.153 (0.762)
I/sI	8.1 (1.7)	9.2 (1.1)
Completeness (%)	94.3 (72.8)	99.4 (96.5)
Redundancy	3.6 (3.3)	3.7 (3.3)
Refinement		
Resolution (Å)	49.3-3.5	48.9-3.3
No. reflections	81,638	105636
R_{work}/ R_{free}	20.1/23.2	20.7 /25
No. atoms		
Protein/Nucleic acid	29180	29151
Ligand/Ions	9	42
Water		
B-factors		
Protein/Nucleic acid	94.3	102.8
Ligand/Ions	135.8	95.4
Water		
R.m.s deviations		
Bond lengths (Å)	0.009	0.009
Bond angles (º)	1.355	1.326

* Values in parentheses are for the highest-resolution shell.

CORRIGENDUM

doi:10.1038/nature14607

Corrigendum: Divergent reprogramming routes lead to alternative stem-cell states

Peter D. Tonge, Andrew J. Corso, Claudio Monetti, Samer M. I. Hussein, Mira C. Puri, Iacovos P. Michael, Mira Li, Dong-Sung Lee, Jessica C. Mar, Nicole Cloonan, David L. Wood, Maely E. Gauthier, Othmar Korn, Jennifer L. Clancy, Thomas Preiss, Sean M. Grimmond, Jong-Yeon Shin, Jeong-Sun Seo, Christine A. Wells, Ian M. Rogers & Andras Nagy

Nature **516,** 192–197 (2014); doi:10.1038/nature14047

In this Article, the address listed as Nicole Cloonan's present address (QIMR Berghofer Medical Research Institute, Queensland 4006, Australia) should have been listed as her other affiliation with superscript 16, because the work she did was split equally between the two institutions. This has been corrected in the online versions of the paper.

RETRACTION

doi:10.1038/nature14591

Retraction: Integrative genomics identifies APOE ε4 effectors in Alzheimer's disease

Herve Rhinn, Ryousuke Fujita, Liang Qiang, Rong Chen, Joseph H. Lee & Asa Abeliovich

Nature **500,** 45–50 (2013); doi:10.1038/nature12415

In this Article, we described integrative genomics analyses of Alzheimer's disease and associated risk factors. However, reanalysis of the data has showed that sample numbers, image panels and data points were inappropriately manipulated and inaccurate in the ELISA and subcellular localization studies presented in Figs 2d, e, 3b, g, h and 4c, as well as in corresponding Supplementary Figs 10–16. We are in the process of repeating these cell-based studies. We remain confident in the transcriptomics and human genetics analyses reported in the Article. However, given these issues, we wish to retract the Article in its entirety. We deeply regret this circumstance and apologize to the community.

CORRIGENDUM

doi:10.1038/nature14606

Corrigendum: Genome-wide characterization of the routes to pluripotency

Samer M. I. Hussein, Mira C. Puri, Peter D. Tonge, Marco Benevento, Andrew J. Corso, Jennifer L. Clancy, Rowland Mosbergen, Mira Li, Dong-Sung Lee, Nicole Cloonan, David L. A.Wood, Javier Munoz, Robert Middleton, Othmar Korn, Hardip R. Patel, Carl A. White, Jong-Yeon Shin, Maely E. Gauthier, Kim-Anh Lê Cao, Jong-Il Kim, Jessica C. Mar, Nika Shakiba, William Ritchie, John E. J. Rasko, Sean M. Grimmond, Peter W. Zandstra, Christine A. Wells, Thomas Preiss, Jeong-Sun Seo, Albert J. R. Heck, Ian M. Rogers & Andras Nagy

Nature **516,** 198–206 (2014); doi:10.1038/nature14046

In this Article, the address listed as Nicole Cloonan's present address (QIMR Berghofer Medical Research Institute, Queensland 4006, Australia) should have been listed as her other affiliation with superscript 22, because the work she did was split equally between the two institutions. This has been corrected in the online versions of the paper.

FIELDWORK An ecologist's story of banking seeds and building skills **p.629**

HOW TO LEAVE ACADEMIA Six tips for gaining transferable skills **go.nature.com/24scaj**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

TINBEE/SHUTTERSTOCK

EXTRAMURAL WORK

To serve or not to serve

When committees come knocking, scientists need to know which requests will benefit them and which will only steal their time — and how to tell the difference.

BY ROBERTA KWOK

Anastasia Ailamaki fondly remembers her first experience serving on a grant-application review committee for the US National Science Foundation (NSF). Through working with peers to evaluate and rank grant proposals asking for spectrometers and other instruments, Ailamaki, a computer scientist now at the Swiss Federal Institute of Technology in Lausanne, gained valuable insight into what makes an application clear and convincing. "I adored that experience," she says. She credits it with helping her to prepare her own successful application for an NSF early-career-development grant.

But like many researchers, Ailamaki has at times been overloaded with requests for her service. "First reaction is that I'm very flattered that I have been invited," she says. "Second is that I realize I really don't have time, by any possible measure, to be on that committee. And the third reaction is to say yes." She has served on committees of all types, including those dealing with promotions, department management, campus events and conference and workshop organization. Although many of these experiences have proved valuable, she now tries to consider requests more carefully before accepting them — weighing, for instance, whether she is uniquely qualified for the spot or whether the committee chair could easily find someone else.

Committee work is tricky for scientists to navigate. On the one hand, it can offer many benefits: opportunities to network, learn about the state of the field, get ideas to improve research and influence funding decisions or policy. On the other hand, some researchers become overburdened — they sacrifice research time to sit in meetings, they draft recommendations that go unused or they get dragged into political disputes. And institutions may lack concrete guidelines for service requirements, making it difficult for researchers to gauge whether their workload is fair.

But careful strategizing can help scientists to make the most of their service. They should gather information about committees before agreeing to join, consider the work's potential impact and proactively seek assignments that they feel passionate about. To help committees to run smoothly, members should actively aim to keep discussions on topic and treat peers respectfully. And as leaders, committee chairs should ensure that the process is efficient and professional (see 'The ruling of the chair').

Junior researchers might feel obligated to accept every committee request. At some institutions, women or researchers from ▶

under-represented minorities, in particular, may be recruited more often than their peers to increase diversity on a panel, and so might feel pressure to serve as a representative voice. But before deciding, scientists should consider whether the assignment is worthwhile for them personally. "You've got to get something out of it as well," says Patricia Molina, head of the physiology department at Louisiana State University Health Sciences Center New Orleans. She also chairs the National Hispanic Science Network, a virtual organization that promotes research on issues important to the Hispanic community and fosters development of Hispanic scientists.

THE POWER TO SAY NO

It can be hard to work out which invitations to turn down because service requirements are sometimes vague and guidelines vary by institution. A regional university with a limited graduate programme might expect faculty members to be heavily involved in university governance — for example, developing policies that are related to undergraduate education — whereas a research-focused university might value service with national and international professional associations.

Researchers should ask their department heads, mentors or colleagues for advice on how to evaluate a request. Senior faculty members might know how much work a committee entails and the extent to which it will benefit a scientist's career. They might also warn of political landmines, such as two departments that fight constantly over the same resources. For instance, a curriculum committee could be time-consuming because of a knotty battle to change entrenched teaching methods, says Maryrose Franko, senior science programme manager at the Howard Hughes Medical Institute's Janelia Research Campus in Ashburn, Virginia.

Scientists should also investigate the potential impact of the group's work. People are often eager to serve on committees that advise federal government agencies because the invitation makes them feel important, says Tom Cech, a biochemist at the University of Colorado Boulder. But he adds that they should ask the chair about the fate of their findings. In some cases, the agency is committed to funding the recommendations, but in others, reports are simply circulated to political staffers with no guarantee that anyone will attempt to implement the ideas.

> *"If we don't get involved, decisions will be made in the absence of scientific input."*

For some scientists, the chance to influence important issues might be worth the risk of wasting time. In 2011–13, geophysicist Steve Hickman served on a committee that advised the US Department of the Interior on improving safety of offshore development of oil and gas. Hickman, who now directs the US Geological Survey Earthquake Science Center in Menlo Park, California, did not know whether the group's advice would be followed. "It is a gamble," he says. "But if we don't get involved, decisions will be made in the absence of scientific input." Their work paid off — some of the group's recommendations, such as setting up an ocean energy-safety institute, are now in place.

Service can also pay off in networking opportunities. Members of a department seminar committee, for example, have a chance to invite speakers in their field whom they would like to meet. These visitors might give the scientist feedback on ongoing projects or write reference letters in the future. Serving with a professional association could enable graduate students and postdocs to meet potential employers, and organizing a conference will earn a researcher name recognition in the field. In 2011, when Megan Carey organized an international neuroscience symposium at her institute, she became acquainted with many of the speakers she had invited — some of whom later asked her to give talks. "It was an incredible networking opportunity for me," says Carey, a neuroscientist at the Champalimaud Centre for the Unknown in Lisbon, Portugal.

And some committee members forge personal, not just professional, connections. When Hickman chaired a science-advisory group for the International Continental Scientific Drilling Program, the team took trips to drilling sites around the world together, which helped to build camaraderie. "Some of my best friends I've made in my field have been on committees like this," he says.

COMMITTEE PHOBIA

For scientists who loathe committees and simply want to do their research, service assignments that benefit their immediate working environment may be the most palatable. By participating in faculty searches, for instance, researchers can select colleagues who could positively influence their work. "Being able to shape your environment is something that's important for all, even for the person who says, 'I just want to get my science done,'" says Jeremy Boss, an immunologist at the Emory University School of Medicine in Atlanta, Georgia. A new colleague could suggest ideas to improve research, such as studies to read or experimental techniques to try.

Researchers may also volunteer for committees that appeal to them, instead of waiting for requests. "The worst thing is to get assigned

EASE THE PAIN

The ruling of the chair

Institutions and committee chairs can take steps to make service assignments fair and painless for researchers. University departments, for instance, can be transparent about how much service work each faculty member is performing, says Joya Misra, a sociologist at the University of Massachusetts Amherst. She says that departments could e-mail researchers every year with all service assigned over the past decade. If some scientists realize that they have been doing more than their colleagues, they might feel less guilty about declining requests. At the University of Maryland in College Park, higher-education researcher KerryAnn O'Meara and her colleagues analysed annual faculty reports to calculate the average number of service activities performed by professors at a given rank and college — for example, an associate professor in the college of computer, mathematical and natural sciences — and published the data on an internal website. Faculty members use the site to compare their service workload to their peers' and decide whether to accept assignments, says O'Meara. She is willing to share templates with other universities to show them how to collect and present similar data.

If researchers bear an unusually heavy service load, institutions can compensate by reducing their teaching requirements, suggests Misra. Creating default rotations for time-intensive departmental roles can also help to distribute the work fairly between faculty members.

Committee chairs should keep the process efficient. When ecologist Jay Stachowicz chaired an educational-policy committee at the University of California, Davis, he began each meeting by noting items that seemed uncontroversial — such as eliminating a course from a major requirement because it was no longer offered — and asking whether anyone opposed passing them. If not, he moved on.

The chair should also ensure that members treat each other respectfully. Senior faculty members who try to bulldoze junior researchers may need a private reminder that each person's opinion counts equally, regardless of his or her rank. "I want everybody to agree that they're going to park their titles at the door," says John Murry Jr, coordinator of the higher-education programme at the University of Arkansas in Fayetteville. If some people are dominating the discussion, the chair can intervene or pull quieter members aside during a break to encourage participation. **R.K.**

to some random committee that you have no passion for," says Cech. Once they have chosen committees for themselves, scientists can use those service obligations as reasons to decline less-desirable assignments.

After committing to a group, scientists should execute their duties diligently — it is always possible that the committee chair will evaluate them for a promotion later.

If the committee's goal is vague or discussions are unfocused, researchers can ask the chair to clarify the mission with administrators or to provide agendas in advance. During meetings, members should avoid making comments that do not directly serve the committee's purpose. For instance, when developing policy, people often tell anecdotes to show why the regulation is necessary, says Boss. "All it does is waste time," he says. Instead, the team should concentrate on the wording of the policy and ensure that it covers the necessary scenarios.

Researchers outside traditional universities may encounter a wide variety of expectations and styles. Scientists at the Janelia Research Campus have minimal service obligations so that they can focus on research, whereas those at the Wilderness Society, a conservation organization in Washington DC, are encouraged to serve on committees that influence policy and management decisions. At the Champalimaud Centre, a small group of neuroscientists has been shaping the direction of the budding programme. Faculty members are involved in more types of service than are those in academia, and their meetings can be more intense and efficient. For example, they all participate in hiring decisions, but rather than interviewing candidates over several months, they gather for a one- or two-day symposium to see applicants give talks.

Scientists should discuss committee-service expectations during their job-offer negotiations. A supervisor might even be able to provide precise requirements. Molina expects junior researchers in her department to spend no more than 5% of their time on committee work; mid-level researchers are expected to spend 10–15%.

Ultimately, science cannot run without service. Researchers need to review each other's proposals, contribute to professional organizations and help universities to foster strong research and student development. Faculty members who avoid all committees risk isolating themselves from the community or being perceived as slackers. "In science, people are expected to be givers and sharers," says Molina. Still, that is no reason to feel guilty for setting boundaries. "I believe in participating and volunteering," she says, "but there's a limit." ■

Roberta Kwok *is a freelance writer in Seattle, Washington.*

CHRISTOPHER TRUE

TURNING POINT Heather Schneider

For her postdoc, ecologist Heather Schneider joined Project Baseline, a nationwide US initiative that is developing a seed bank for future scientists to study how plants are evolving in response to climate change. The project has left her little time for her own research at the University of California, Santa Barbara, but the skills she has gained have broadened her career avenues.

What is a field season like?

It's really daunting. Project Baseline's goal is to collect seeds from 43 species — at 10 sites for each one. The project so far has collected 3 million seeds from species both native and introduced. My adviser, Susan Mazer, and I oversee collection in the western region — 237 distinct plant populations of 20 species — and this is the final of 3 field seasons. I spend January to March getting field permits to collect specimens in national and state parks, nature preserves and the University of California reserves. Then I use herbarium records to find historical populations. I try to visit each of our sites twice a season — once while plants are in bloom, to find populations more easily and to collect environmental data, and again to gather seeds. Last year, our field season ended in mid-October.

What about this project lured you away from a pure research focus?

Few things are as important as understanding how ecosystems will respond to climate change. I was interested in helping to create a resource that would be useful for both basic and applied science for the next 50 years. To me, that would have a big impact on ecology and evolutionary biology — much bigger than any single paper I would ever write. I also felt that I have the set of skills — field botany, plant identification and collection of herbarium specimens — necessary for the job.

Did it feel risky to move away from conventional research?

A little. Although my career trajectory has zigzagged, there has been one underlying theme — assessing the impact of human-made threats to ecosystems. I have focused on invasive species, air pollution and habitat degradation. I joke that when you work on short-term grants, you end up with a long tail of 'publications in progress' that follow you from job to job. I'm still working on papers from one to two jobs ago. So it was appealing that there would be less pressure to publish in this position, which could give me a chance to catch up on papers I'm still working on.

Does publishing less concern you?

The principal investigators on the project made sure that our efforts benefited my and the other postdocs' careers. Susan and I work on a greenhouse experiment in the off-season, when we're not in the field for Project Baseline. We have one paper in revision and one in review, so I still am getting papers out.

What are your hopes for future use of this resource?

The research possibilities are huge. Given my own interests, I hope that people will use it to look at ecological interactions. For example, as pollinator communities change, how will that affect wild-plant reproduction? I'm also interested in what the weedy species will do — will the geographical areas where they are found shrink or expand?

What are your job prospects?

I would be interested in a teaching job at a smaller university. I am OK not ending up at a top-tier research university because funding rates are not that encouraging. And the skills I have gained on Project Baseline — project management, budgets, organization, troubleshooting — are applicable to all kinds of other jobs.

Do you plan to promote use of Project Baseline data in future?

Yes. The postdocs on the project want to feel that this resource will be well cared for. I know there are plans to advertise it widely. The principal investigators invited all the postdocs to be on the advisory board, and it is nice to know that we will have a part in evaluating the proposals for its use in the future. ■

INTERVIEW BY VIRGINIA GEWIN

This interview has been edited for length and clarity.

THE SHOULDER OF ORION

A life-changing experience.

BY ERIC GARSIDE

Ric Williamson was lying on the grass and watching the stars fly by overhead. Judging by the constellations, he estimated that they were about halfway between Saturn and Uranus.

"Ric, query," announced the small grey-skinned being lying next to him.

"Go on then, Norbert."

"What will happen to my consciousness; after I am dead?"

Ric grinned, laced his fingers together and put his hands behind his head.

"Well, nothing. That's sort of what dead means, Norbert."

"Have I done something to upset you? Do you no longer have a need of me?"

Ric laughed a bit.

"Oh we do, Norbert. We will always have need of you. In fact, you'll start being reprinted once the *Gaia* exits the gate."

The smile faded from Ric's face. This wouldn't be the first time he'd prepared a Synthetic to die, but that didn't make breaking your friend's digital heart any easier.

"How many times have I died, Ric?"

"Eleven."

Waves of sadness and confusion crashed upon the silicon shores of Norbert's mind.

"I don't want to die, Ric."

"Hold out your hand," Ric said, reaching deep into the soil and grabbing a handful of earth. "Do you know what that contains? Death. Bits of dead star, of dead planet; traces of plants, fragments of animals. All things die, Norbert. But in their death, new life takes root."

Norbert used his fingers to push the pile of dirt around in his palm.

"Our lives and deaths rest in your hands, Norbert. We *need* you to stay behind and destroy the jumpgate once the *Gaia* is through. If the Black Mass finds out where we've gone, they will follow."

Without breaking his gaze from the dirt, Norbert replied simply: "I understand."

Norbert climbed the ladder to the cockpit of his starfighter, with Ric close behind. As Norbert strapped himself in, Ric reached into his pocket and fished something out of it. Dangling from a small plastic cord was an oddly shaped bit of metal, which he placed in Norbert's hand.

"It's a good luck charm."

"But you told me luck was just people taking probability personally, Ric."

ILLUSTRATION BY JACEY

Norbert's stoic gaze met one last sly wink from Ric before the flight shield secured in place.

"Think of me sometimes, Norbert," Ric yelled, his voice muffled almost entirely by the shield.

The *Gaia* lurched painfully slowly towards the jump gate. Beams of pure chaos erupted in every direction as the gate powered up. A splash of brilliant light, and the moon-sized mothership was gone. As the engine trails of the *Gaia* faded into the dark, the onboard computer in Norbert's craft snapped to life.

PRIMARY OBJECTIVE COMPLETE; THE GAIA *HAS LEFT THE SECTOR.*

NEW PRIMARY OBJECTIVE ENABLED: DESTROY THE TANNHAUSER GATE.

The weapons system of the craft enabled, and Norbert set to work. The gate was massive. Simply to fly round its circumference took almost an hour, and destroying it was very systematic work. First the shield generators needed to be taken out. Then the back-up reactors, followed by the primary reactors, and finally the data centres.

Explosion after explosion tallied Norbert's progress, each pass over the gate leaving less of it than before. After long gruelling hours of work, the final data centre erupted in a quick burst of flame.

PRIMARY OBJECTIVE COMPLETE; THE TANNHAUSER GATE IS DISABLED

TABULA RASA ENGAGED.

Norbert looked towards the computer, his face contorting in distress. The flight systems of the craft were disabled, and a new waypoint was assigned at the centre of the gate. Fear so dominated Norbert's emotion core that it automatically shut down to prevent an overload.

REACTION LIMITERS DISABLED; OVERLOAD IN APPROXIMATELY 5 MINUTES.

His emotion subsystem came back online just in time for him to appreciate the crushing realization of his own mortality. Panicked, the synthetic humanoid frantically tried to activate any control to avert his imminent demise. Every button played a denial sound, every controller moved freely without responsive feedback.

If mechanical men could cry, Norbert certainly would have done. Instead, he was forced to come to grips with the knowledge of his death simply by hanging his head forlornly.

That's when he noticed it, hanging by a plastic thread on his chest; the odd bit of metal Ric had given him. He took it in his hands, running his fingers over the channels and grooves in the amulet. It reminded him of…

A key! A spark of realization hit, and Norbert reached under his seat and pulled back the floor panelling to expose the manual-override panel. He slotted the key in place, closed his eyes and turned it. A small metal click sounded. Norbert tossed away the panel face.

Within the chamber was a rather large, rather red button surrounded by caution tape and labelled "Thermonuclear Engine Ejection". Norbert smashed the button with incredible force, and a comically small reactor fired backwards out of the ship.

With the ship running on reserve power, all non-essentials (such as automated control) were disabled, returning command of the craft to Norbert.

His chances of survival were slim, there was no doubting that. He might have enough power to outrun the explosion, but did he have enough to get anywhere after that? And to what end? And for how long?

And that's when he realized it; the meaning of life.

It was wanting to live.

In this moment, a perfect storm of chaos and clarity overcame him. Norbert became more than wires, diodes and synthetic emotions; he became truly human.

Grasping the controls firmly, he fixed his gaze on the stars, and took his chance. ■

Eric Garside *is an educational software developer with a passion for science, technology and astronomy.*